IMPLEMENTASI ALGORITMA *DECISION TREE* UNTUK MENDETEKSI *MULTI-LABEL HATE SPEECH* DAN *ABUSIVE LANGUAGE* PADA TWITTER BAHASA INDONESIA

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Teknik
Pada Jurusan Teknik Informatika

Oleh :

FAUZI IHSAN
11651100236

UIN SUSKA RIAU

FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU
PEKANBARU
2021

LEMBAR PERSETUJUAN

IMPLEMENTASI ALGORITMA *DECISION TREE* UNTUK MENDETEKSI *MULTI-LABEL HATE SPEECH* DAN *ABUSIVE LANGUAGE* PADA TWITTER BAHASA INDONESIA

TUGAS AKHIR

Oleh

FAUZI IHSAN
11651100236

Telah diperiksa dan disetujui sebagai Laporan Tugas Akhir

di Pekanbaru, pada tanggal 20 Januari 2021

Pembimbing

Surya Agustian, S.T., M.Kom.
NIP. 19760803 201101 1 003

LEMBAR PENGESAHAN

IMPLEMENTASI ALGORITMA *DECISION TREE* UNTUK MENDETEKSI *MULTI-LABEL HATE SPEECH* DAN *ABUSIVE LANGUAGE* PADA TWITTER BAHASA INDONESIA

TUGAS AKHIR

Oleh

FAUZI IHSAN
11651100236

Telah dipertahankan di depan sidang dewan penguji
sebagai salah satu syarat untuk memperoleh gelar sarjana Teknik Informatika
Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau
di Pekanbaru, pada tanggal 20 Januari 2021

Pekanbaru, 20 Januari 2021
Mengesahkan

Dekan	**Ketua Jurusan**
Dr. Drs. Ahmad Darmawi, M.Ag.	**Dr. Elin Haerani, S.T., M.Kom.**
NIP. 19660604 199203 1 004	**NIP. 19810523 200710 2 003**

DEWAN PENGUJI

Ketua	: Muhammad Fikry, S.T., M.Sc.
Sekretaris	: Surya Agustian, S.T., M.Kom.
Penguji I	: Iwan Iskandar, M.T.
Penguji II	: Nazruddin Safaat H., M.T.

iii

LEMBAR HAK ATAS KEKAYAAN INTELEKTUAL

Tugas Akhir yang tidak diterbitkan ini terdaftar dan tersedia di Perpustakaan Universitas Islam Negeri Sultan Syarif Kasim Riau adalah terbuka untuk umum dengan ketentuan bahwa hak cipta pada penulis. Referensi kepustakaan diperkenankan dicatat, tetapi pengutipan atau ringkasan hanya dapat dilakukan seizin penulis dan harus disertai dengan kebiasaan ilmiah untuk menyebutkan sumbernya.

Penggandaan atau penerbitan sebagian atau seluruh Tugas Akhir ini harus memperoleh izin dari Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau. Perpustakaan yang meminjamkan Tugas Akhir ini untuk anggotanya diharapkan untuk mengisi nama, tanda peminjaman dan tanggal pinjam.

LEMBAR PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan didalam daftar pustaka.

Pekanbaru, 20 Januari 2021

Yang membuat pernyataan,

FAUZI IHSAN
11651100236

LEMBAR PERSEMBAHAN

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

يَٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُوٓاْ إِذَا قِيلَ لَكُمْ تَفَسَّحُواْ فِى ٱلْمَجَٰلِسِ فَٱفْسَحُواْ يَفْسَحِ ٱللَّهُ لَكُمْ ۖ وَإِذَا قِيلَ ٱنشُزُواْ فَٱنشُزُواْ يَرْفَعِ ٱللَّهُ ٱلَّذِينَ ءَامَنُواْ مِنكُمْ وَٱلَّذِينَ أُوتُواْ ٱلْعِلْمَ دَرَجَٰتٍ ۚ وَٱللَّهُ بِمَا تَعْمَلُونَ خَبِيرٌ ﴿١١﴾

Alhamdulillah...

Segala Puji dan Syukur kepada-Mu Yaa Allah yang maha kaya akan Ilmu karena-Mu lah hamba dapat menyelesaikan Tugas Akhir ini...

***

Tugas Akhir ini kupersembahan untuk kedua orang tua, yang telah sangat berjasa, memberikan segala dukungan, semangat dan doa yang tak kenal lelah yang tidak bisa dinilai dengan angka untuk anakmu ini. Terima kasih Orang Tua ku...

***

Dan kupersembahkan untuk kakak, abang, adik-adik, dan keponakan ku tersayang, terima kasih dukungan dan doanya yang telah diberikan selama ini...

***

Dan juga kepada keluarga dan kerabat, yang selalu memberikan motivasi dikala jenuh dan kesulitan dalam hidup. Karena kalianlah aku mampu untuk berjuang sekeras mungkin hingga mencapai titik ini...

Alhamdulillah, Allah telah menganugerahiku Keluarga yang indah...

IMPLEMENTASI ALGORITMA *DECISION TREE* UNTUK MENDETEKSI *MULTI-LABEL HATE SPEECH* DAN *ABUSIVE LANGUAGE* PADA TWITTER BAHASA INDONESIA

FAUZI IHSAN
11651100236

Tanggal Sidang : 20 Januari 2021
Periode Wisuda :

Jurusan Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Sultan Syarif Kasim Riau

ABSTRAK

Ujaran kebencian dan bahasa kasar mudah ditemukan di dalam komunikasi tertulis di social media seperti twitter, yang dapat memicu terjadinya persengketaan di antara korban dan pengujarnya. Penyaringan yang dilakukan dengan manual tidak dapat terukur dan membutuhkan waktu yang lama. Bagaimanapun, akan sulit memeriksa apakah suatu tweet mengandung ujaran kebencian dan/atau Bahasa kasar bila seseorang berpihak. Penelitian ini bertujuan untuk membangun pemodelan untuk mengklasifikasi tweet apakah mengandung ujaran kebencian dan kata-kata kasar. Apabila terdeteksi mengandung ujaran kebencian, maka level ujaran kebenciannya diukur. Dataset yang digunakan terdiri dari 13.126 cuitan asli twitter. Metode *Decision Tree* digunakan sebagai metode Klasifikasi berdasarkan hasil pengujian metode *Decision Tree* menggunakan *Feature Engineering* berupa fitur khusus, tekstual dan *lexicon* serta tanpa penggunaan kombinasi *preprocessing* seperti *case folding*, *stopword removal* dan *punctuation removal* mendapatkan akurasi rata-rata tertinggi sebesar **71,03%** dengan nilai akurasi *hate speech* sebesar **71,52%**, nilai akurasi *abusive* sebesar **78,07%** dan nilai akurasi level sebesar **63,49%**.

Kata Kunci : Bahasa Kasar, *Decision Tree*, *Feature Engineering*, Klasifikasi, *Twitter*, Ujaran Kebencian

IMPLEMENTATION OF DECISION TREE ALGORITHM TO DETECT MULTI-LABEL HATE SPEECH AND ABUSIVE LANGUAGE IN INDONESIAN TWITTER

FAUZI IHSAN
11651100236

Date of Final Exam : 20th Januari 2021
Date of Graduation Ceremony :

Informatic Engineering Department
Faculty of Science and Technology
State Islamic University of Sultan Syarif Kasim Riau

ABSTRACT

Hate speech and abusive language are easily found in written communications in social media like twitter. They often generate a dispute between both parties, the victims and the first who write the tweet. Filtering tweets manually is immeasurable and takes a long time. However, it is almost difficult to distinguish a tweet contains hate speech and/or abusive language, for ones who take sides. This research is to build a modelling to classify the tweets into class of abusive and/or contain hate speech. If hate speech detected, then the hardness level of hatred will be measured. The dataset contains 13,126 real tweets data. Decision Tree method is used as a classification method. Based on the test results of the Decision Tree method using Feature Engineering in the form of special, textual and lexicon features and without the use of preprocessing combinations such as case folding, stopword removal and punctuation removal, the highest average accuracy is 71.03% with an accuracy value of hate speech 71.52%, the abusive accuracy value is 78.07% and the accuracy value of level is 63.49%.

Keywords : *Abusive Language, Decision Tree, Feature Engineering, Classification, Twitter, Hate Speech*

KATA PENGANTAR

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Assalamua'alaikum Warahmatullahi Wabarakatuh

Alhamdulillahirobbil'aalamiin. Puji syukur kepada Allah SWT atas seluruh rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan penelitian dan penulisan Laporan Tugas Akhir di Jurusan Teknik Informatika dengan Judul "Implementasi Algoritma *Decision Tree* untuk mendeteksi *Multi-Label Hate Speech* dan *Abusive Language* pada Twitter Bahasa Indonesia". Laporan Tugas Akhir ini merupakan prasyarat untuk memperoleh gelar sarjana strata satu di Jurusan Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim Riau (UIN SUSKA RIAU).

Selama dilaksanakannya penelitian dan penulisan Laporan Tugas Akhir ini, penulis mendapatkan banyak pengetahuan, pengalaman, bimbingan, dukungan dan juga arahan dari semua pihak yang telah membantu hingga Laporan Tugas Akhir ini dapat diselesaikan. Untuk itu pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada:

1. Allah SWT dengan segala rahmat-Nya memberikan semua yang terbaik dan dengan hidayah-Nya memberikan petunjuk sehingga dalam penyusunan laporan ini berjalan lancar.
2. Bapak Prof. Dr. Suyitno, M.Ag., selaku Plt. Rektor Universitas Islam Negeri Sultan Syarif Kasim Riau.
3. Bapak Dr. Drs. Ahmad Darmawi, M.Ag., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau.
4. Ibu Dr. Elin Haerani, S.T., M.Kom., selaku Ketua Jurusan Teknik Informatika Universitas Islam Negeri Sultan Syarif Kasim Riau.

5. Bapak Surya Agustian, S.T, M.Kom., selaku dosen pembimbing Tugas Akhir penulis, yang telah sangat banyak berbagi waktu, ilmu dan wawasan yang dimiliki kepada penulis sehingga penelitian dan Laporan Tugas Akhir ini dapat diselesaikan.
6. Bapak Iwan Iskandar, M.T., selaku dosen penguji I yang telah meluangkan waktunya dan banyak memberikan saran, dan masukan yang membangun bagi penulis.
7. Bapak Nazruddin Safaat Harahap, M.T., selaku dosen penguji II yang telah meluangkan waktunya dan banyak memberikan saran dan masukan yang membangun bagi penulis.
8. Bapak Febi Yanto, M.Kom., selaku Pembimbing Akademis penulis selama menjalani perkuliahan di Jurusan Teknik Informatika, yang telah banyak memberikan arahan serta masukan mengenai perkuliahan bagi penulis.
9. Seluruh Bapak/Ibu dosen Teknik Informatika yang telah sabar memberikan tunjuk ajar serta ilmu yang bermanfaat kepada penulis selama masa perkuliahan.
10. Kedua orang tua penulis, Bah dan Mak yang selalu memberi dukungan dan doa yang tiada henti. Terima kasih atas semangat dan nasehat yang sangat tulus kepada anaknya. Semoga Allah SWT senantiasa memberikan kesehatan kepada Bah dan Mak.
11. Kakak, Abang dan Adik (Ummi Farida, Jhon Hendri, Fitri Dewi, Robi Sugiarto, Napis Muzamil, Fatimah Aszaira, Afrison Azurda) yang juga memberikan semangat, nasehat, dan doa yang tulus untuk saudaranya. Semoga Allah SWT senantiasa memberikan kesehatan kepada kalian semua.
12. Keempat keponakanku (M. Ziyad Zul Iqbal, Syifa Elsya Ramadhani, Muhammad Dzaka Abdillah dan Bunga Salsabila) yang menjadi penghibur penulis dikampung halaman dan menjadi penyemangat penulis untuk menyelesaikan tugas akhir ini. Semoga Allah selalu memberikan kesehatan kepada kalian nak.
13. Teman-teman seperjuangan Teknik Informatika kelas B angkatan 2016 yang selalu memberikan semangat untuk terus berjuang.

14. Teman-teman Kos Awak, yang telah banyak memberikan saran, semangat, dan motivasi kepada penulis, semoga kita semua dapat mencapai apa yang diinginkan.
15. Teman-teman SMA yang selalu mendukung, mendoakan dan memberikan semangat. Semoga kita sukses kawan-kawan dan tercapai semua keinginan.
16. Semua pihak yang tidak dapat penulis sebutkan satu persatu. Terimakasih atas dukungan moril maupun materil dalam pengerjaan Laporan Tugas Akhir ini.

Semoga laporan ini dapat bermanfaat bagi penulis khususnya maupun pembaca pada umumnya. Serta dapat menjadi referensi dan rujukan bagi hal-hal yang bermanfaat. Penulis berharap adanya kritik dan saran guna memperbaiki atau sebagai pengembangan kedepannya. Akhir kata penulis ucapkan terima kasih dan selamat membaca.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Pekanbaru, 20 Januari 2021

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Ujaran kebencian (*Hate Speech*) merupakan suatu ungkapan langsung maupun tidak langsung yang menuju kepada individu ataupun kepada suatu kelompok yang mengandung kebencian berdasarkan suatu hal yang melekat pada individu ataupun kelompok tersebut yang menyerang agama, etnis, *gender*, dan orientasi seksual. Ujaran kebencian merupakan sesuatu perkataan, sikap, tulisan maupun sesuatu pertunjukan yang dilarang sebab bisa merangsang munculnya aksi kekerasan serta perilaku prasangka, baik itu dari pihak pelakon yang membagikan pernyataan tersebut maupun korban dari aksi tersebut (Febriyani, 2018).

Dalam kehidupan sehari-hari, media sosial menjadi tempat suatu individu ataupun kelompok dalam melakukan penyebaran ujaran kebencian dan sering juga disertai dengan bahasa kasar (*Abusive Language*) (Davidson, Warmsley, Macy, & Weber, 2017). Bahasa kasar umumnya diucapkan serta dituliskan lalu digunakan sebagai serangan kepada pihak tertentu, tujuannya untuk menyampaikan kekesalan, kekecewaan maupun meluapkan emosi tentang kejadian tertentu. Pengungkapan kata kasar bisa diungkapkan dengan mengatakan kategori hewan tertentu, semacam anjing, monyet serta lainnya. Tetapi tidak seluruh kalimat yang memuat kategori hewan termasuk kedalam bahasa kasar (Hidayatullah, Yusuf, Juwairi, & Nayoan, 2019)

Media sosial membuat komunikasi cepat tersampaikan, hal tersebut merupakan hal positif yang kita dapatkan dari penggunaan media sosial, namun bukan berarti kita bisa dengan bebas menggunakannya sesuai dengan keinginan kita. Media sosial juga memiliki peraturan yang diatur oleh negara agar tidak terjadi penyalahgunaan media sosial yang dapat merugikan orang lain. Meski telah diatur

oleh negara, tak sedikit juga orang-orang yang tetap menyalahgunakan media sosial untuk kepentingan pribadi.

Bentuk penyalahgunaan media social diantaranya, penipuan, penyebaran ujaran kebencian (*hate specch*), bahasa kasar (*abusive language*), dan lainnya. Hal tersebut mudah saja terjadi kepada siapapun, oleh sebab itu, perlunya pengetahuan mengenai penggunaan media sosial agar dapat terhindar dari orang-orang yang menyalahgunakan media sosial untuk kepentingan pribadi.

Pencemaran nama baik atau fitnah dan perilaku tidak menyenangkan ialah perbuatan melawan hukum sebab melanggar hak orang lain. Kalaupun sikap tersebut tidak berlangsung secara langsung dan terbuka, sikap tersebut umumnya dilakukan di dunia maya yaitu pada media sosial, karena di media sosial, orang dapat dengan bebas mengemukakan pendapat atau mengkritik seseorang atau sekelompok orang, dan karena tidak ada kontak fisik maka dia merasa tidak melanggar hukum yang berlaku.

Contoh pelanggaran di media sosial adalah seorang netizen bernama Zikri Dzatil yang mengunggah postingan di media sosial tentang penghinaan terhadap Walikota Surabaya Rismaharini. Netizen tersebut menggunggah foto Rismaharini tengah mengatur lalu lintas saat banjir di Surabaya dan menyebut Risma sebagai "Kodok Betina". Postingan tersebut dianggap pencemaran nama baik sehingga Zikria diamankan Satreskrim Polrestabes Surabaya di kota Bogor pada 31 januari 2020 dan akibat tindakannya itu Zikria terancam Pasal 45A Ayat (2) jo Pasal 28 Ayat (2) UU Nomor 19 Tahun 2016 tentang Perubahan UU Nomor 11 Tahun 2008 tentang ITE serta Pasal 45 Ayat (3) jo Pasal 27 Ayat (3) UU 19 Tahun 2016 tentang Perubahan UU Nomor 11 Tahun 2008 tentang ITE.[1]

Ada beberapa pendapat terkait postingan Zikria Dzatil. Meskipun Zikria beralasan karena banyaknya netizen yang membandingkan kinerja Gubernur Anies Baswedan dan Walikota Risma dalam menangani banjir, namun sebagian

[1] https://surabaya.kompas.com/read/2020/02/06/11570491/sakit-hati-anies-di-bully-soal-banjir-jakarta-motif-zikria-dzatil-hina-risma?page=all

masyarakat masih menganggap postingan tersebut tidak pantas dan tergolong pencemaran nama baik karena secara langsung menyasar pada suatu objek. Oleh sebab itu, perlu dilakukan penelitian lebih lanjut tentang ujaran kebencian, jenis-jenis kalimat yang masuk kedalam ujaran kebencian, dan tingkat ujaran kebencian, agar masyarakat dapat membedakan mana kalimat ujaran kebencian atau yang bukan ujaran kebencian.

Dibanding dengan media sosial lain, kelebihan *Twitter* menurut (Putra, 2014) antara lain jangkauannya sangat luas, tidak cuma teman, namun juga ada publik figur, potensi periklanan lebih besar kedepannya, serta komunikasi terjalin dengan cepat (*update*), *mulitilink* (terhubung kepada orang banyak) dan lebih terukur daripada *Facebook. Twitter* dapat menyebarkan informasi secara cepat, yang bakal menjadi topik diskusi di antara para penggunanya. Media massa, seperti koran, tabloid, majalah dan televisi juga memanfaatkan *Twitter* guna menyebarkan berita. Karena dengan *Twitter* media massa setiap saat bisa *mengupdate* berita melalui *Twitter*, masyarakat bisa lebih mudah mendapatkan informasi dan *update* dengan cepat.

Beberapa kajian yang dilakukan sebelumnya mengenai ujaran kebencian, salah satunya penelitian yang dilakukan oleh (Hakiem & Fauzi, 2019) dengan menggunakan metode *Naive Bayes* berbasis *N-Gram* dengan Seleksi Fitur *Information Gain.* Penelitian ini bertujuan untuk mengklasifikasikan *tweet* kedalam ujaran kebencian atau bukan ujaran kebencian. Data yang digunakan merupakan data *tweet* berbahasa indonesia sebanyak 500 data yang terdiri dari 250 data dengan label ujaran kebencian, dan 250 data dengan label bukan ujaran kebencian. Fitur *N-Gram* yang digunakan pada penelitian ini, yaitu fitur *Unigram, Bigram* dan kombinasi antara *Unigram-Bigram.* Perbandingan data latih dan data uji adalah 80% untuk data latih dan 20% untuk data uji. Hasil akurasi terbaik didapat pada fitur *Unigram* tanpa menggunakan seleksi fitur *Information Gain,* yaitu sebesar 84%, nilai *precision* sebesar 92%, nilai *recall* sebesar 79,31%, dan nilai *f-measure* sebesar 85,18%. Dari penelitian ini didapatkan kesimpulan bahwa hasil akurasi dari klasifikasi ujaran kebencian pada twitter dengan menggunakan motode *naive bayes*

mendapatkan hasil paling akurat dengan menggunakan fitur *unigram* dan tanpa melakukan seleksi fitur *information gain.*

Pada penelitan yang diteliti oleh (Luqyana, Cholissodin, & Perdana, 2018) tentang Analisis Sentimen *Cyberbullying* pada Komentar Instagram dengan Metode Klasifikasi *Support Vector Machine*. Penelitian ini bertujuan untuk mengklasifikasikan data instagram tentang *cyberbullying* menjadi 2, yaitu sentimen positif dan negatif. Analisi ini dilakukan karena *cyberbullying* sangat berbahaya, dan tentunya juga meresahkan banyak orang akibat dampak yang ditimbulkan. Data yang digunakan sebanyak 400 data yang diambil secara offline dengan total fitur 1799, pembobotan yang dilakukan yaitu menggunakan *Term Frequency-Inverse Document Frequency* (*TF-IDF*) dan menggunakan metode klasifikasi *SVM*. Perbandingan data latih dan data uji yaitu 70% banding 30%. Berdasarkan pengujian, didapatkan hasil parameter terbaik algoritma *SVM* ialah pada nilai *degree kernel polynomial* sebesar 2, nilai *learning rate* sebesar 0,0001, dan iterasi maksimum sebanyak 200 kali. Hasil pengujian tersebut mendapatkan akurasi tertinggi pada pembagian data latih 50% dan data uji 50% yaitu sebesar 90%.

Berkaitan dengan penelitian yang akan peneliti ambil menggunakan metode *decision tree*, terdapat beberapa penelitian terkait diantaranya, penelitian yang dilakukan oleh (Romadloni, Santoso, & Budilaksono, 2019) tentang perbandingan metode *naive bayes, knn dan decision tree* terhadap analisis sentimen transportasi krl *commnuter line.* Data twitter yang digunakan sebanyak 127 data, diproses kedalam beberapa tahapan, yaitu *convert emoticon, cleansing, case folding, tokenizing,* dan *stemming.* Hasil uji coba pada metode *naive bayes* mendapatkan akurasi sebesar 80%, *precision* sebesar 66,67%, *sensitivity* sebesar 100% dan *specificity* sebesar 66,67%. Pada metode *KNN* mendapatkan akurasi sebesar 80%, *precision* sebesar 100%, *sensitivity* sebesar 50% dan *specificity* sebesar 100%, sedangkan pada metode *decision tree* mendapatkan akurasi sebesar 100%, *precision* sebesar 100%, *sensitivity* sebesar 100% dan *specificity* sebesar 100%

Berdasarkan paparan latar belakang diatas, maka pada penelitian ini penulis akan merancang dan membangun sebuah model yang dapat mengklasifikasikan

multi-label hate speech dan *abusive languange* pada twitter Bahasa Indonesia menggunakan metode *Decision Tree.*

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini yaitu bagaimana cara mengklasifikasikan *multi-label* dan *multi-class hate speech* pada twitter bahasa indonesia mengunakan algoritma *decision tree* dan menghitung akurasi dari algoritma *decision tree* dalam mengklasifikasikan *hate speech* pada twitter berbahasa indonesia.

1.3 Batasan Masalah

Batasan masalah pada penelitian ini antara lain:

1. *Dataset* yang digunakan sebanyak 13.126 *tweet* (Ibrohim & Budi, 2019)
2. Label yang diproses yaitu *hate speech, abusive* dan level
3. Level *hate speech* yang dideteksi yaitu level *strong hate speech, medium hate speech* dan *weak hate speech*
4. Metode *Decision Tree* yang digunakan untuk klasifikasi adalah metode *Classification and Regression Tree*
5. Metode *Embedding* kata yang digunakan adalah *FastText*
6. Hasil keluaran yang diharapkan dari penelitian ini yaitu klasifikasi *multi-class* dan *multi-label hate speech* dan *abusive language* pada twitter bahasa indonesia

1.4 Tujuan Penelitian

Tujuan dari penelitian tugas akhir ini adalah:

1. Menerapkan algoritma *decision tree* untuk mengklasifikasikan *multi-class* dan *multi-label hate speech* dan *abusive language* pada twitter berbahasa indonesia
2. Menghitung nilai akurasi dari algoritma *decision tree* dalam pengklasifikasian *hate speech* pada *tweet* berbahasa indonesia

3. Melakukan proses *feature engineering* untuk menemukan *set feature* terbaik untuk mendapatkan model yang mampu memprediksi dengan lebih akurat

1.5 Sistematika Penulisan

Sistematika penulisan penelitian ini terdiri dari 6 bab,yaitu sebagai berikut:

BAB I PENDAHULUAN

Pada bab pendahuluan berisi latar belakang dari pemasalahan yang penulis angkat, rumusan masalah berdasarkan latar belakang, batasan masalah agar penelitian tidak keluar dari topik pembahasan, tujuan penelitian serta sistematika penulisan pada penelitian tugas akhir ini.

BAB II LANDASAN TEORI

Pada bab landasan teori, peneliti membahas teori yang mendukung untuk proses pengerjaan tugas akhir mengenai *hate speech, twitter,* algoritam *decision tree* dan penelitian terdahulu yang terkait.

BAB III METODOLOGI PENELITIAN

Pada bab metodologi penelitian akan membahas tahapan penelitian yang dilakukan pada tugas akhir ini.

BAB IV ANALISA DAN PERANCANGAN

Pada bab ini membahas mengenai analisa data dan untuk membuat pemodelan sesuai dengan tujuan penelitian dan perancangan pemodelan.

BAB V IMPLEMENTASI DAN PENGUJIAN

Pada bab implementasi dan pengujian akan mengimplementasikan dan melakukan pengujian pada pemodelan yang dibangun sesuai

dengan hasil analisa dan perancangan sebelumnya, dan melakukan pengujian terhadap pemodelan.

BAB VI PENUTUP

Bab penutup berisi kesimpulan serta saran dari penelitian tugas akhir yang telah diteliti sebelumnya.

BAB II
LANDASAN TEORI

2.1 Twiiter

Twitter didirikan oleh Jack Dorsey pada 21 Maret 2006. Sejak rilis pertama pada bulan Juli, *Twitter* menjadi situs yang paling sering dikunjungi diinternet. Jumlah penggunanya semakin meningkat, dengan 300.000 pengguna perhari (Buntoro, 2017). Pengguna *twitter* dimungkinkan dapat membaca dan mengirim teks mencapai 140 karakter, namun pada tanggal 7 November 2017 meningkat menjadi 280 karakter, sehingga disebut dengan *tweet*.

Twitter merupakan layanan dalam bentuk jejaring sosial dan pengguna Internet membutuhkannya sebagai media komunikasi untuk memperoleh informasi. Informasi di *Twitter* bisa berupa opini, bisa juga pertanyaan ataupun berupa komentar positif dan komentar negatif (Nurjanah, Perdana, & Fauzi, 2017). Lebih dari 500 juta *tweet* dikirim oleh *user* setiap hari dan *Twitter* mendapat lebih dari 1,6 miliar pencarian untuk setiap harinya. Beberapa istilah yang sering digunakan pada *Twitter*, antara lain *Direct Message, Follow, Favorite, Follower, Unfollow, Tweet, Following, Retweet, Hashtag, Mention, Timeline, Trending Topic, List, Search, Over Heard (OH).*

2.2 *Hate Speech*

Ujaran kebencian merupakan tindakan berupa komunikasi dan dilakukan oleh individu maupun kelompok berupa provokasi, hasutan, atau hinaan menuju kepada individu dan kelompok lainnya yang berkaitan dengan aspek seperti warna kulit, etnis, ras, gender, orientasi seksual, kewarganegaraan, cacat, agama, dan lainnya.

Ujaran kebencian bisa dilakukan melalui bermacam cara seperti melalui orasi pada kegiatan kampanye, spanduk, serta media sosial, penyampaiannya dapat dilakukandimuka umum seperti demonstrasi, pada kegiatan keagamaan, maupun pada media masa, baik media cetak maupun media elektronik, dan kadang melalui pamflet. (Febriyani, 2018)

Ancaman pidana juga ditujukan bagi setiap orang dengan sengaja dan tanpa hak menyebarkan informasi yang ditujukan untuk menimbulkan rasa kebencian atau permusuhan individu atau kelompok masyarakat tertentu berdasarkan atas SARA (Pasal 28 ayat (1) dan Pasal 45 UU ITE). Tindak pidana ini juga dirumuskan secara materil. Artinya, tindak pidana selesai sempurna akibat adanya rasa kebencian atau permusuhan antar kelompok masyarakat telah timbul.

Selain UU ITE, Pasal 207 dan Pasal 310-Pasal 321 Kitab Undang-Undang Hukum Pidana (KUHP) juga memuat larangan melakukan penghinaan, dengan segala bentuknya, yang menyerang kehormatan dan nama baik. Substansi dalam pasal-pasal ini telah dimuat kembali dalam RUU KUHP. Dalam RUU KUHP, yang dimaksud dengan "penghinaan" adalah menyerang kehormatan atau nama baik orang lain. Sifat dari perbuatan pencemaran adalah jika perbuatan penghinaan yang dilakukan dengan cara menuduh, baik secara lisan, tulisan, maupun dengan gambar yang menyerang kehormatan dan nama baik seseorang, sehingga merugikan orang tersebut.

2.3 *Abusive Language* (Bahasa Kasar)

Penyebaran ujaran kebencian di media sosial sering disertai dengan bahasa kasar (Davidson et al., 2017). Bahasa kasar adalah ucapan yang mengandung kata-kata kasar/frasa yang disampaikan kepada lawan bicaranya (individu atau kelompok), baik secara verbal maupun secara tertulis. Di Indonesia, kata-kata kasar biasanya berasal dari kondisi yang tidak menyenangkan seperti gangguan mental, penyimpangan seksual, cacat fisik ataupun nama-nama hewan yang memiliki karakteristik buruk dan menjijikkan, dan terlarang dalam agama tertentu. Namun tidak semua kalimat yang memuat jenis hewan anggap kedalam bahasa kasar (Hidayatullah et al., 2019)

2.4 *Text Mining*

Menurut (Feldman, 2008), konsep *text mining* hampir sama dengan konsep pada *data mining* biasanya, namun pada *text* mining data yang digunakan ialah data teks, jadi pada *text mining* akan menganalisa teks tersebut untuk menemukan pola

pada data teks. Biasanya digunakan untuk masalah analisa sentimen, klasifikasi, pengelompokan teks serta ekstraksi informasi dan pengambilan informasi.

T*ext mining* memiliki beberapa tahapan pengolahan utama, yaitu pemrosesan awal *text* atau disebut dengan *text preprocessing*, *text transformation*, *feature selection* dan yang terakhir adalah penemuan (*pattern discovery*) (Srivastava, A. N., & Sahami, 2009)

a. *Text*

Permasalahan yang biasa pada penerapan *text mining* hampir sama halnya dengan *data mining*, seperti data dalam jumlah yang sangat besar, dimensi data yang sangat tinggi, data dan strukturnya yang dinamis serta *data noise*. Data yang digunakan pada *text mining* biasanya digunakan data tidak terstruktur dan semi terstruktur sedangkan pada *data mining* data yang digunakan merupakan data terstruktur.

b. *Text Preprocessing*

Karena data yang tidak terstruktur maka pada tahap ini data diolah agar bisa dipergunakan untuk tahap selanjutnya, biasanya data teks dipecah dan dibagi-bagi menjadi beberapa bagian serta mengubah huruf menjadi huruf besar atau kecil. Selain itu juga dilakukan penghapusan tanda baca dan angka-angka. (Feldman, 2008).

c. *Text Transformation* (*Feature Generation*)

Text Transformation bisa juga disebut dengan *Fitur Extraction*, dimana kita akan mendapat representasi dari data teks tersebut kedalam bentuk array yang berisi angka-angka tertentu dan dengan dimensi tertentu, teknik yang biasa digunakan seperti *bag of word, tf-idf* dan yang sedang populer yaitu *word embeddings.*

d. *Feature Selection*

Menurut (Nugroho & Wibowo, 2017) *feature selection* ialah salah satu teknik yang digunakan untuk menentukan atribut pada data yang paling berpengaruh, semakin tinggi dimensi dari data, maka akan semakin memperlambat proses komputasi, sedangkan tidak semua atribut bernilai

penting dan berpengaruh, disinilah peran dari seleksi fitur, yaitu untuk menyeleksi atribut yang berpengaruh atau tidak, agar bisa dikurangi untuk mempercepat proses komputasi nantinya.

e. *Pattern discovery*

Tahap penemuan pola merupakan tahap yang paling penting, disinilah peran dari teknik-teknik data mining yang akan digunakan untuk mengekstraksi pola yang terdapat pada teks, seperti melakukan klasifikasi pada teks ataupun melakukan clustering pada data teks.

f. *Interpretation/evaluation*

Interpretation/evaluation merupakan proses untuk menampilkan hasil dari tahap penemuan pola agar mudah dimengerti oleh pembaca yang ditampilkan kedalam bentuk tertentu misalnya kedalam visual seperti tabel dan grafik.

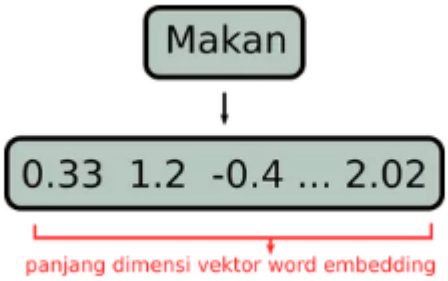
2.5 *Word Embedding*

Word embedding merupakan suatu teknik yang digunakan untuk mempresentasikan sebuah kata menjadi sebuah *vector* dengan dimensi tertentu. *Word embedding* merupakan pengembangan komputasi pemodelan kata-kata yang sederhana seperti perhitungan menggunakan jumlah dan frekuensi kemunculan kata dalam sebuah dokumen. *Word embedding* menggambarkan kedekatan sebuah kata atau dokumen dalam kontekstual sesuai dengan data latih yang digunakan dalam pembentukannya, sehingga seringkali kedekatan tersebut bukan merupakan makna dari sebuah kata.

Dalam *word embedding* dapat digambarkan bahwa setiap "kata" diwakilkan oleh sebuah titik di dalam luasan bidang tertentu, titik-titik ini kemudian akan dipelajari oleh perhitungan *word embedding* dan satu titik akan dipindahkan menjauh atau mendekati titik yang lainnya, berdasarkan kata-kata lain yang mengelilingi titik tersebut. Hal ini dilakukan terus menerus hingga sampai pada sebuah kondisi dimana semua titik tidak dapat dipindahkan lagi mendekati (atau menjauhi) titik yang lainnya. Sehingga hasil akhir dari iterasi ini dapat memberikan sebuah gambaran dimana kata-kata dengan makna yang serupa akan cenderung

berada dalam satu area yang sama dalam bidang tersebut atau dengan kata lain kata-kata yang ada dalam satu area pada bidang tersebut dan mempunyai jarak kedekatan yang kecil cenderung mempunyai kesamaan.

Metode *Word embedding* mengkonversi "kata" menjadi "vektor" yang berisi angka-angka dengan ukuran yang cukup kecil untuk mengandung informasi yang lebih banyak. Informasi yang didapat akan cukup banyak hingga vektor dapat mendeteksi makna, misal kata "marah" dan "mengamuk" itu lebih memiliki kedekatan nilai dibandingkan kata "marah" dan "bahagia". Illustrasi dari *word embedding* dapat diliha pada gambar dibawah ini.



Gambar 2. 1 Ilustrasi *Word Embedding*

Penelitian ini menggunakan metode *word embedding* yang diberi nama *FastText* yang dibuat oleh Facebook AI Research (FAIR). *FastText* adalah perpustakaan untuk mempelajari *embeddings* kata dan klasifikasi teks. Model ini memungkinkan untuk membuat pembelajaran tanpa pengawasan atau algoritma pembelajaran terawasi untuk mendapatkan representasi *vector* untuk kata-kata. *Facebook* menyediakan model pra-pelatihan untuk 294 bahasa, termasuk bahasa Indonesia. *FastText* menggunakan jaringan saraf untuk penyematan kata.

Penggunaan panjang dimensi *vector* dalam proses pembuatan model biasanya tergantung dari data yang digunakan. Pada penelitian yang dilakukan oleh (Alfariqi, Maharani, & Husen, 2020) mereka melakukan penelitian menggunakan 3.582 data twitter dari hasil *crawling* dan menggunakan panjang dimensi *vector* berukuran 100 guna untuk meningkatkan performansi model. Hal ini juga mempertimbangkan perangkat keras yang digunakan dan *cost* yang diperlukan dalam proses *pre-trained*

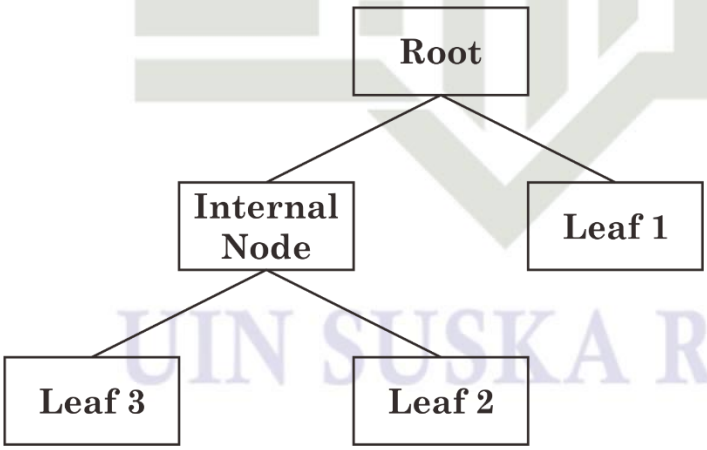
FastText. Pada penelitian yang dilakukan oleh (Arliyanti Nurdin, Bernadus Anggo Seno Aji, Anugrayani Bustamin, 2020) mereka menggunakan data pada korpus Wikipedia, UMBC dan berita *statmt.org* lalu menggunakan 300 dimensi *vector* kata.

2.6 *Decision Tree*

Decision Tree ialah sebuah pohon terbalik yang dimulai dari akar dan setiap cabang menampilkan sejumlah pilihan yang ada, sedangkan daun merupakan keputusan akhir yang dipilih. Keunggulan *Decision Tree* yaitu daerah pengambilan keputusan yang sebelumnya sangan kompleks dan luas, dapat diringkas menjadi lebih spesifik dan simpel. Metode ini dapat mengurangi jumlah kriteria pada setiap node internal tanpa mengurangi hasil kualitas keputusannya. (Setiawati, Taufik, & Z, 2016)

2.6.1 Struktur Dasar

Secara umum, algoritma *decision tree* dimulai dari sebuah titik awal (*root*). Dari titik awal ini, pengguna akan memecah sesuai dengan algoritma *decision tree*. Hasilnya adalah sebuah pohon keputusan dimana kemungkinan skenario yang dihasilkan ditunjukkan oleh setiap cabangnya serta hasil dari keputusan pada cabng tersebut.



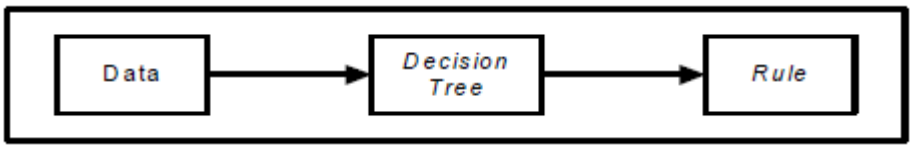
Gambar 2. 2 Konsep dasar pohon keputusan

Pohon keputusan terdiri dari *root*, *internal node* merupakan pembagian berdasarkan hasil uji, sedangkan *leaf* merupakan kelas yang dihasilkan.

2.6.2 Proses pengembangan *Decision Tree*

Algoritma *decision tree* sangat populer dan banyak digunakan secara praktis. Algoritma pohon keputusan berusaha menemukan fungsi pendekatan yang bernilai diskrit dan tahap terhadap kesalahan data. (Ariestya, Praptiningsih, & Supriatin, 2016)

Konsep pohon keputusan yaitu mengubah data yang ada pada tabel keputusan menjadi sebuah pohon keputusan dan aturan-aturan keputusan. Berikut adalah konsep dari *decision tree*



Gambar 2. 3 Konsep *decision tree*

2.6.3 Langkah-langkah algoritma *decision tree*

Secara umum langkah-langkah untuk membangun sebuah pohon keputusan adalah sebagai berikut:

a. Menentukan node terpilih

   Menentukan node yang terpilih menggunakan nilai *Entropy* dari setiap kriteria dengan data sampel yang ditentukan dan node yang terpilih adalah kriteria dengan *Entropy* yang paling kecil.

   Untuk menghitung nilai *Entropy* adalah sebagai berikut:

$$Entropy(s) = \sum_{i=l}^{n} -pi * log_2\, pi$$

b. Membuat pohon keputusan

   Periksa apakah nilai *entropy* dari anggota Node ada yang bernilai nol. Jika ada, tentukan daun yang terbentuk. Jika seluruh nilai *entropy* anggota Node adalah nol, maka proses pun berhenti. Jika ada anggota Node yang memiliki nilai *entropy* lebih besar dari nol, ulangi lagi proses dari awal dengan Node sebagai syarat sampai semua anggota dari Node bernilai nol.

c. Mengubah *tree* menjadi *rule*

Setelah pohon keputusan dibuat, maka ubah menjadi sebuah *rule* menggunakan IF-THEN.

2.7 *Feature Engineering*

Feature Engineering (Rekayasa Fitur) adalah suatu cara yang dilakukan untuk memilih dan membuat fitur sendiri dengan menggunakan pengetahuan agar model klasifikasi *machine learning* bisa bekerja lebih akurat. Ini melibatkan pemilihan fitur atau subset informatif dan/atau kombinasi fitur berbeda menjadi fitur baru untuk mendapatkan representasi klasifikasi. Dalam domain klasifikasi teks, biasanya fitur menyertakan istilah yang berbeda dalam corpus. Bahkan dalam korpus kecil dapat memiliki banyak sekali fitur, berpotensi membutuhkan *feature engineering* untuk klasifikasi. Pengetahuan domain sering digunakan untuk memandu *feature engineering*, contohnya mendefinisikan kamus istilah yang terkait dengan penyakit X (Garla & Brandt, 2012)

2.8 Python

Python ialah bahasa pemrograman multifungsi yang diciptakan oleh Guido van Rossum dan dirilis pada tahun 1991. GvR, begitu ia biasa disebut di komunitas Python, menciptakan Python untuk menjadi interpreter yang memiliki kemampuan penanganan kesalahan (*exception handling*) dan mengutamakan sintaksis yang mudah dibaca serta dimengerti (*readability*). Didesain untuk memudahkan dalam prototyping, Python menjadi bahasa yang sangat mudah dipahami dan fleksibel.

Python juga memilih untuk menggunakan indentasi untuk mengelompokkan blok kode, berbeda dengan beberapa bahasa lain yang menggunakan simbol tertentu, misalnya kurung kurawal, atau sintaksis begin-end. Sehingga secara visual pun, blok kode Python didesain untuk mudah dipahami. Salah satu yang paling dikenal adalah, penggunaan titik koma atau semicolon (;) tidak wajib di Python dan penggunaan semicolon cenderung dianggap bukan cara khas Python (*non-pythonic way*), meskipun ia tetap dapat digunakan, misalnya untuk memisahkan dua statement dalam baris yang sama.

2.9 *Confusion Matrix*

Confusion matrix merupakan cara yang digunakan untuk menentukan apakah objek yang dideteksi tersebut benar atau salah (Hastuti, 2012). Biasa nilai yang dicari menggunakan *confusion matrix* yaitu nilai akurasi, nilai presisi dan nilai recall. (Baihaqi, Handayani, & Pujianto, 2019). Tabel berikut merupakan penjelasan mengenai *confusion matrix*.

Tabel 2. 1 Tabel *Confusion Matrix*

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif
Positif	TP (*True Positive*)	FN (*False Positive*)
Negatif	FN (*False Negative*)	TN (*True Negative*)

Akurasi = $\frac{(TP+TN)}{TP+TN+FP+FN} x100$

Precision = $\frac{TP}{TP+FP} x100$

Recall = $\frac{TP}{TP+FN} x100$

Keterangan:

TP : *True Positive* adalah data kelas positif yang diprediksi sistem kedalam kelas positif

TN : *True Negative* adalah data kelas negatif yang diprediksi sistem kedalam kelas negatif

FP : *False Positive* adalah data kelas positif yang diprediksi sistem kedalam kelas negatif

FN : *False Negative* adalah data kelas negatif yang diprediksi sistem kedalam kelas positif

2.10 Penelitian Terkait

Pada tabel 2. 2 berikut merupakan penelitian/kajian yang pernah diteliti yang berkaitan dengan penelitan tugas akhir ini.

Tabel 2. 2 Penelitian Terkait

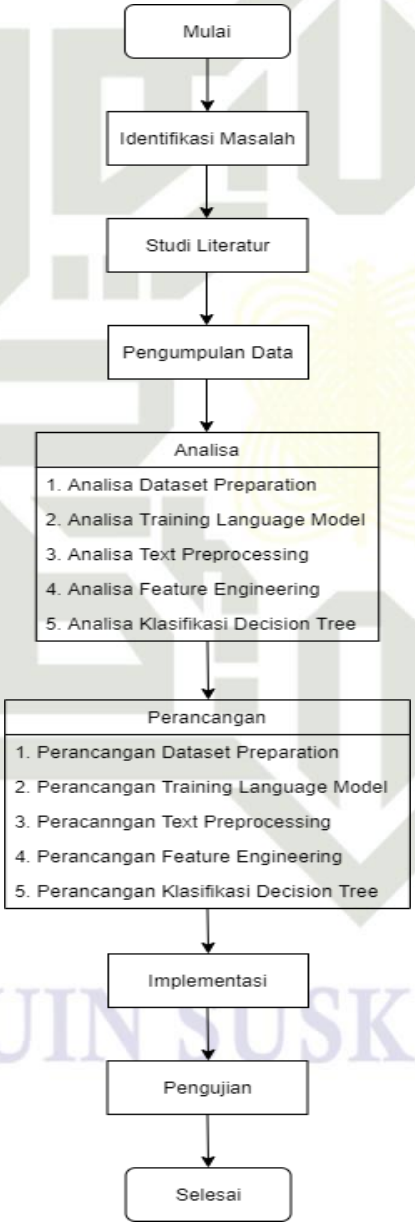
No	*Author*	Tahun	Judul Penelitian	Hasil
1	Nurjanah, dkk	2017	Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter Menggunakan Metode *K-Nearest Neighbor* dan Pembobotan Jumlah *Retweet*	Akurasi menggunakan pembobontan teks 82,50%, akurasi menggunakan pembobotan non-tekstual 60%, dan kombinasi keduanya 83,33% dengan nilai *k*=3 dan konstanta perkalian α=0,8 dan β=0,2
2	Rofiqoh, dkk	2017	Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan *Lexicon Based Features*	Hasil akurasi sistem sebesar 79% dengan nilai *degree* = 2, *learning rate* 0,0001, jumlah iterasi maksimum 50 kali. Sedangkan tanpa memakai *Lexicon Based Features* akurasi sebesar 84% dengan nilai kesamaan parameter.
3	Imelda A.Muis, Muhammad Affandes, M.T	2015	Penerapan Metode *Support Vector Machine* (SVM) Menggunakan Kernel *Radial Basis Function* (RBF) Pada Klasifikasi *Tweet*	Data yang tidak dilakukan pemilihan fitur akurasi = 97,54%, sedangkan untuk data yang dilakukan pemilihan fitur nilai akurasi = 99,12%.
4	Romadloni, dkk	2019	Perbandingan metode naive bayes, knn dan decision Tree terhadap analisis sentimen transportasi krl *Commuter line*	Metode Bayes mendapat akurasi = 80%, precision = 66,67%, sensitivity = 100%, specificity = 66,67%. Metode KNN akurasi = 80%, precision = 100%, sensitivity = 50%, specificity = 100% dan Metode Decision Tree akurasi = 100%, precision = 100%, sensitivity = 100%, specificity = 100%
5	Yusra, dkk	2016	Perbandingan Klasifikasi Tugas Akhir Mahasiswa	Naïve Bayes mendapat nilai terbaik = 87%.

No	*Author*	Tahun	Judul Penelitian	Hasil
			Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor	KNN mendapat nilai akurasi 84% dengan nilai k=3, 85% dengan nilai k=5, 86% dengan nilai k=7 dan 84% dengan nilai k=9
6	Agustina Merdeka Raya, dkk	2019	Klasifikasi Sentimen Masyarakat Terhadap Kenaikan Harga Tiket Pesawat Pada Twitter Menggunakan Naive Bayes	Akurasi bayes = 90,70% sedangkan akurasi KNN = sebesar 67,79%
7	Ghulam Asrofi Buntoro	2017	Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter	Naïve Bayes Classifier mendapat nilai tertinggi dengan nilai rata-rata akurasi 95%, nilai presisi 95%, nilai recall 95% nilai TP rate 96,8% dan nilai TN rate 84,6%.
8	Ghulam Asrofi Buntoro	2016	Analisis Sentimen *Hate Speech* pada Twitter dengan Metode *Naive Bayes Classifier* dan *Support Vector Machine*	Akurasi terbaik pada metode *support vector machine* memakai tokenisasi unigram dan stopword *list* Bahasa Indonesia serta *emoticons*, dengan rata-rata akurasi = 66,6%, presisi = 67,1%,*recall* = 66,7% *TP rate* = 66,7% dan *TN rate* = 75,8%.
9	Meri Febriyani	2018	Analisis faktor penyebab pelaku melakukan ujaran kebencian (*hate speech*) dalam media sosial	Faktor yang menyebabkan pelaku mengutarakan ujaran kebencian di media sosial, yaitu faktor-faktor dari dalam diri individu (internal) termasuk psikologi individu dan keadaan kejiwaan dan faktor diluar individu yaitu faktor lingkungan, faktor kekurangan kontrol sosial, faktor kepentingan umum, faktor ketidaktahuan

No	*Author*	Tahun	Judul Penelitian	Hasil
				komunitas, faktor fasilitas dan faktor kemajuan teknologi.
10	Muhammad Hakiem, dkk	2019	Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur *Information Gain*	Memakai fitur *Unigram*, *Bigram*, dan kombinasi *unigram-bigram*. Data yang digunakan sebanyak 250 data berlabel ujaran kebencian dan 250 data berlabel bukan ujaran kebencian dengan perbandingan 80%:20%. Akurasi terbaik menggunakan fitur *Unigram* dan tanpa menggunakan seleksi fitur Information Gain. Hasil akurasi terbaik 84%, nilai *precision* 92%, nilai *recall* 79,31%, dan nilai *f-measure* 85,18%.
11	Nurdifa Febrianti, dkk	2019	Klasifikasi *Tweet* Berbahasa Indonesia Berisi Ujaran Kebencian Menggunakan Metode *Improved K-Nearest Neighbor* dengan Pembobotan BM25F	Pembobotan nilai stream pada BM25F mempengaruhi hasil klasifikasi IKNN. Sedangkan hasil akhir terbaik untuk *F-Measure*, *Accuracy*, *Precision*, dan *Recall* dari rerata *5-Fold Cross Validation* yang didapatkan ialah 79,77% , 68,80%, 68,80%, dan 89,92% dengan k = 70, bs = 0,6, $v1$ = 2, $v2$ = 5 dan $k1$ = 2 sebagai nilai terbaik untuk setiap parameternya.
12	Shiddiq, dkk	2018	Analisa Kepuasan Konsumen Menggunakan Klasifikasi *Decision Tree* di Restoran Dapur Solo (Cabang Kediri)	Berdasarkan hasil record dari 300 data. Didapatkan analsisa pelanggan yang puas yaitu sebanyak 93,9%

BAB III
METODOLOGI PENELITIAN

Metodologi penelitian merupakan tahapan penelitian. Metodologi penelitian memiliki tujuan untuk mendapatkan hasil yang diharapkan peneliti. Berikut tahapan penelitian tugas akhir ini:



Gambar 3. 1 Tahapan Metodologi Penelitian

3.1 Identifikasi Masalah

Pada tahap identifikasi masalah berisikan latar belakang masalah, rumusan masalah penelitian, serta batasan masalah penelitian. Landasan penelitian akan dicantumkan pada latar belakang, masalah yang dihadapi berdasarkan latar belakang dituangkan dirumusan masalah, batasan masalah berperan untuk membatasi agar penelitian tidak melebar kearah lain, dan harapan atau tujuan dituangkan pada tujuan penelitian.

3.2 Studi Literatur

Studi literatur merupakan tahap untuk memperoleh semua informasi melalui jurnal, buku, *paper* internasional dan referensi lainnya yang sangat berhubungan dengan penelitian ini seperti teori-teori tentang penelitian yang pernah dilakukan sebelumnya.

3.3 Pengumpulan Data

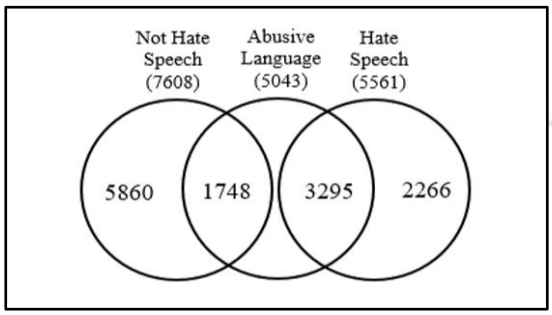
Data yang diperlukan pada penelitian ini adalah data twitter berupa *tweet* yang memiliki kemungkinan termasuk ke dalam *tweet hate speech* dan *abusive*. Proses pengumpulan data tidak dilakukan dengan cara memilih dan memberi label pada data *tweet* secara manual, karena pada penelitian sebelumnya yang dilakukan oleh (Ibrohim & Budi, 2019) tentang *hate speech* dan *abusive language* telah melakukan *crawling* data *tweet hate speech* dan *abusive*.

(Ibrohim & Budi, 2019) mempersilakan siapa saja yang ingin melakukan penelitian dengan topik yang berhubungan dan memerlukan dataset tentang *hate speech* dan *abusive* dapat menggunakan dataset yang sama secara gratis (*Free*). Dalam penelitian ini, menggunakan *dataset* kebencian dan penyalahgunaan bahasa Twitter dari beberapa penelitian sebelumnya yang terdiri dari (Ibrohim & Budi, 2019). Selain menggunakan *dataset* Twitter dari penelitian sebelumnya, (Ibrohim & Budi, 2019) juga merangkak *tweet* untuk memperkaya *dataset* sehingga dapat mencakup jenis penulisan pidato kebencian dan bahasa kasar yang mungkin belum ada dalam data dari penelitian sebelumnya.

Untuk proses anotasi, pada penelitian sebelumnya membangun sistem anotasi berbasis web untuk memudahkan dalam membuat anotasi data sehingga dapat mempercepat proses anotasi dan meminimalkan kesalahan anotasi. Kami juga melakukan anotasi standar emas untuk menguji apakah bahasa sudah memahami tugas atau tidak. Dalam penelitian ini, kami melakukan diskusi dan konsultasi dengan ahli Bahasa untuk mendapatkan pedoman anotasi yang valid dan anotasi standar emas. Data *Twitter* yang digunakan untuk standar emas berasal dari penelitian sebelumnya dan buku pedoman ucapan kebencian (Komnas HAM, 2015)

Pada tahap anotasi pertama, berhasil mengumpulkan 16.500 *tweet* dari proses perayapan dan penelitian sebelumnya. Dari fase ini, kita mendapatkan 11.292 (68,44% total *tweet* yang dijelaskan dalam fase pertama) yang terdiri dari 6.187 *tweet* ucapan tidak kebencian dan 5.105 *tweet* ucapan benci yang memiliki perjanjian 100%. Selanjutnya, dalam fase anotasi kedua, tercatat 5.700 *tweet* ucapan kebencian (5.105 *tweet* dari anotasi fase pertama dan 595 *tweet* dari (Ibrohim & Budi, 2019)).

Dari proses anotasi dua fase ini, berhasil mendapatkan 13.169 *tweet* yang telah digunakan untuk eksperimen penelitian yang terdiri dari 7.608 *tweet*s ucapan tidak membenci (6.187 *tweet*s dari anotasi tahap pertama dan 1.421 *tweet*s dari (Ibrohim & Budi, 2019)) dan 5.561 *tweet* ucapan kebencian . Distribusi Bahasa kasar ke *tweet*s ucapan tidak membenci pidato dan *tweet*s ucapan benci dari *tweet* yang dikumpulkan dapat dilihat pada Gambar 3.2. Dari Gambar 3.2, kita dapat melihat bahwa tidak semua pidato kebencian adalah Bahasa yang kasar. Sebaliknya, bahasa yang kasar juga tidak harus berupa pidato kebencian.



Gambar 3. 2 Distribusi *Tweet Hate Speech,* Tidak *Hate Speech* dan *Abusive*

3.4 Analisa

Tahap ini akan menjelaskan bagaimana tahapan-tahapan terpenting pada penelitian ini. Tahapan-tahapan tersebut meliputi *dataset preparation*, *feature engineering* dan klasifikasi menggunakan metode *decision tree.*

3.4.1 *Dataset Preparation*

Pada tahap ini dataset yang berasal dari (Ibrohim & Budi, 2019) terdapat 3 label, yaitu *hate speech, abusive,* dan level dari *hatespeech.* Pada label *hate speech* terdapat 2 kelas yaitu *hate speech* dan tidak *hate speech.* Pada label *abusive* terdapat 2 kelas yaitu *abusive* dan tidak *abusive*, sedangkan pada label level terdapat 3 kelas yaitu kuat, sedang dan lemah.

3.4.2 *Training Language Model*

Pada tahap ini akan dilakukan *word embedding* menggunakan library FastText. FastText merupakan *library* yang dikeluarkan oleh *facebook.* FastText merupakan pengembangan dari *library* Word2Vec. FastText mampu menangani kata yang tidak pernah dijumpai sebelumnya. Pada tahap ini setiap kata diubah kedalam *vector* sepanjang 128. Setelah pelatihan selesai dilanjutkan ke tahap berikutnya.

3.4.3 *Text Preprocessing*

Pada tahap ini dilakukan beberapa tahap *preprocessing* seperti *case folding, stopword removal* dan *punctuation removal*. Dari ketiga proses *preprocessing* tersebut akan dilakukan percobaan terhadap kombinasi kemungkinan *preprocessing* yang terbentuk.

3.4.4 *Feature Engineering*

Feature Engineering merupakan suatu proses yang menggunakan pengetahuan kita untuk menyeleksi *features* ataupun membuat *features* baru supaya model *machine learning* dalam memecahkan masalah menjadi akurat. Tahap ini menghabiskan sebagian besar waktu. Jikalau proses ini kita lakukan secara baik, maka model *machine learning* yang dibentuk akan dapat memprediksi dengan lebih tepat dan akurat.

Feature Engineering ialah tahapan yang penting dilakukan dipenelitian ini, karena penulis akan menghabiskan waktu yang cukup lama pada tahap ini untuk mecari *feature* terbaik dari *dateset* yang ada, agar *feature* yang dihasilkan dapat memecahkan masalah dengan lebih akurat.

3.4.5 Klasifikasi *Decision Tree*

Setelah menyelesaikan semua proses, langkah terakhir adalah menggunakan metode *decision tree* untuk klasifikasi. Proses klasifikasi menggunakan parameter terbaik terbaik yang diperoleh sebelumnya dan penggunaan *feature engineering* untuk melakukan proses klasifikasi guna mendapatkan hasil akurasi *decision* terbaik.

3.5 Perancangan

Pada tahap ini pemodelan dirancang berdasarkan hasil dari analisa yang telah dilakukan sebelumnya, hal ini agar mempermudah dalam proses pembuatan model yang terstruktur dengan baik.

Dalam melakukan perancangan menggunakan *tool Jupyter Notebook* sebagai tempat untuk mengimplementasikan kode program python. Tahapan perancangan pada penelitian ini adalah sebagai berikut:

1. Perancangan *Dataset Preparation*

   Pada tahap ini dilakukan penghapusan data *noise* yang tidak diperlukan. Agar data yang digunakan nantinya berkualitas maka perlu menyeleksi terlebih dahulu data yang tidak bagus. Setelah itu baru dapat dilanjutkan kedalam proses berikutnya.
2. Perancangan *Language Model*

   Karena pada saat klasifikasi data yang diolah merupakan data dalam bentuk numerik, maka perlu dilakukan proses *feature extraction* dari teks yang digunakan. Data yang digunakan pada proses *training language model* merupakan data latih saja.

3. Perancangan *Text preprocessing*

   Pada tahap ini akan dilakukan *preprocessing* terhadap teks yang digunakan, meluputi *case folding*, *stopword removal* dan *punctuation removal*.

4. Perancangan *Feature Engineering*

   Pada tahap ini dilakukan perhitungan bobot fitur tambahan terhadap teks yang digunakan, fitur yang dihitung meliputi fitur khusus, tekstual dan *lexicon*.

5. Perancangan Klasifikasi *Decision Tree*

   Pada tahap ini akan dilakukan percobaan model klasifikasi yang sudah dilatih menggunakan *confusion matrix* ialah suatu cara pengujian untuk memperkirakan obyek yang benar dan salah (Hastuti, 2012). Nilai yang diukur biasanya adalah nilai akurasi, presisi dan nilai recall (Baihaqi et al., 2019)

3.6 Implementasi

Implementasi merupakan proses penerapan hasil dari analisa yang dilakukan kedalam sebuah pemodelan. Untuk melakukan implementasi pemodelan dibutuhkan *hardware* dan *software*.

1. *Hardware*:
   a. *Processor* : Intel(R) CoreTM i3 5005U @ 2.0 GHz
   b. *GPU* : Nvidia GEFORCE 930M
   c. *Memory* : 8 GB
   d. *Harddisk* : 500 GB
2. *Software*:
   a. *Operating System* : Windows 10 Profesional
   b. *Programming Language* : Python 3
   c. *Web Browser* : *Chrome*
   d. *Tools* : *Jupyter Notebook*

3.7 Pengujian

Pengujian berisikan implementasi dari analisa dan perancangan pemodelan yang dibangun sesuai tujuan yang ingin dicapai. *Confusion matrix* digunakan untuk mengukur akurasi nantinya.

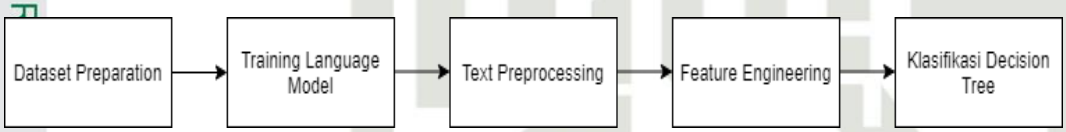
3.8 Kesimpulan dan Saran

Kesimpulan memuat hasil yang telah dicapai pada penelitian, sedankan saran berisikan masukan yang perlu dilakukan untuk penelitian berikutnya.

BAB IV
ANALISA DAN PERANCANGAN

4.1 Analisa

Pada tahap Analisa akan dilakukan pembahasan tentang *dataset preparation, text preprocessing, languange training model, feature engineering* dan klasifikasi algoritma *decision tree* untuk mengetahui hasil akurasi yang diperoleh. Berikut tahapan analisa yang peneliti lakukan.



Gambar 4. 1 Tahapan Penelitian

Penjelasan lebih lanjut mengenai tahapan diatas akan dijelaskan pada penjelasan dibawah ini.

4.1.1 Analisa *Dataset Preparation*

Pada tahap awal yang dikerjakan adalah mempersiapkan *dataset* untuk maju ke tahap berikutnya. *Dataset* didapat dari kajian sebelumnya (Ibrohim & Budi, 2019) dengan jumlah 13.169 data berupa *tweet*. Data ini memiliki 5 label, yaitu HS, *Abusive*, HS Weak, HS Moderate dan HS Strong, namun ada sebanyak 43 data yang *noise*, berupa data kosong dan ada juga yang hanya berisi tag, sehingga data yang digunakan pada penelitian ini akhirnya menjadi 13.126 data. Tabel berikut ini menyajikan dataset awal.

Tabel 4. 1 Dataset Awal

NO	Tweet	Hate Speech	Abusive	Hate Speech Level		
				Weak	Moderate	Strong
1	napadah shopee kampang'	1	1	1	0	0
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0	0	0	1	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1	0	0	0	0
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0	0	0	0	0
5	USER kolam buaya kagak ada qey'	0	0	0	0	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0	0	0	0	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--"'	0	0	0	0	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	1	1	0	1	0
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1	1	1	0	0
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0	0	0	0	0
...	...	...	...	...	...	...
13126	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0	0	0	0	0

Tabel diatas merupakan *dataset* awal pada penelitian ini, tabel diatas menampilkan 10 data awal dan 1 data akhir.

Dataset diatas akan dikelompokkan menjadi 3 kelas, yaitu HS, *Abusive* dan Level. Kelas baru yang terbentuk adalah Level yang merupakan label pada HS_Weak, HS_Moderate dan HS_Strong yang ditransformasikan kedalam Kelas baru dengan nilai 1 untuk HS_Weak, nilai 2 untuk HS_Moderate dan nilai 3 untuk HS_Strong. Setelah dilakukan proses pelabelan Level ini, dataset berubah seperti tabel berikut.

Tabel 4. 2 Dataset Setelah Pelabelan Level

NO	Tweet	HS	Abusive	Hate Speech Level			Level
				Weak	Moderate	Strong	
1	napadah shopee kampang'	1	1	1	0	0	2
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0	0	0	1	0	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1	0	0	0	0	1
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0	0	0	0	0	0
5	USER kolam buaya kagak ada qey'	0	0	0	0	0	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0	0	0	0	0	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--'"	0	0	0	0	0	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	1	1	0	1	0	2
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1	1	1	0	0	1

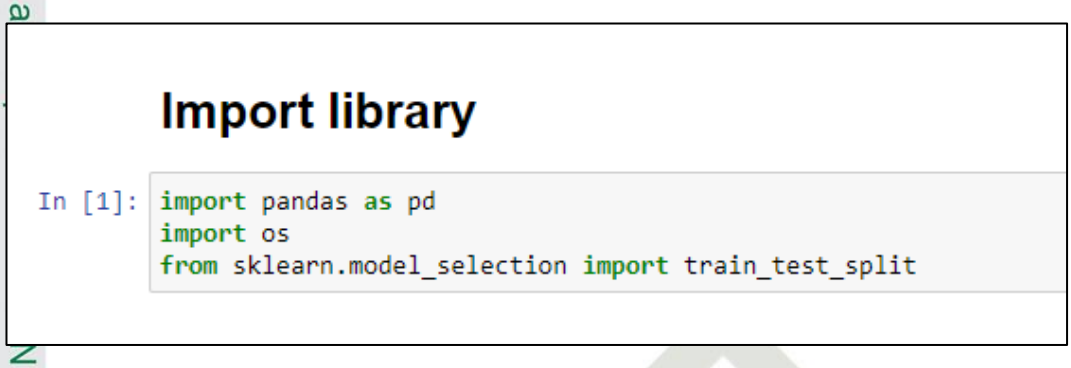
NO	Tweet	HS	Abusive	Hate Speech Level			Level
				Weak	Moderate	Strong	
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0	0	0	0	0	0
...	...	...	...	...	...	...	...
13126	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0	0	0	0	0	0

Pada tahap selanjutnya *dataset* diatas akan dikelompokkan menjadi tiga label, yaitu HS, *Abusive* dan Level. Pada proses awal untuk mempersiapkan data ini menggunakan pemrograman python.

Tahapan yang dilakukan untuk persiapan data ini menggunakan bahasa pemrograman python yaitu:

1. Mengimport *dataset* awal
2. Melakukan Pemisahan label
3. Menyimpan hasil dari pemisahan label
4. Pembagian data untuk data latih dan data uji

Hal pertama yang perlu dilakukan adalah mengimport *library* yang diperlukan. Setelah itu dilanjutkan dengan mengimport datasetnya, seperti pada gambar dibawah.



Gambar 4. 2 *Import library*

Python memiliki banyak sekali *library open source* yang bisa kita gunakan dengan mudah. Seperti *library* pandas berfungsi untuk mengimport data dan dapat menyimpan data kembali kedalam berbagai format seperti csv. Selain itu ada juga *library* os yang salah satu fungsinya untuk membuat directori baru untuk menyimpan file yang diperlukan. Sedangkan *library* sklearn memiliki banyak sekali fungsi yang dalam tahap ini digunakan untuk *splitting* data agar lebih mudah.

Import data

```
In [2]: data = pd.read_csv('Dataset/twitter.csv', ';')
        data
```

Out[2]:

	Tweet	HS	Abusive	HS_Weak	HS_Moderate	HS_Strong	Level
0	- disaat semua cowok berusaha melacak perhatia...	1	1	1	0	0	1
1	RT USER: USER siapa yang telat ngasih tau elu?...	0	1	0	0	0	0
2	41. Kadang aku berfikir, kenapa aku tetap perc...	0	0	0	0	0	0
3	USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...	0	0	0	0	0	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1	1	0	1	0	2
...	...	...	...	...	...	...	...
13121	USER jangan asal ngomong ndasmu. congor lu yg ...	1	1	1	0	0	1
13122	USER Kasur mana enak kunyuk'	0	1	0	0	0	0
13123	USER Hati hati bisu :(.g\n\nlagi bosan huft \...	0	0	0	0	0	0
13124	USER USER USER USER Bom yang real mudah terdet...	0	0	0	0	0	0
13125	USER Mana situ ngasih(": itu cuma foto ya kuti...	1	1	1	0	0	1

13126 rows × 7 columns

Gambar 4. 3 *Import dataset*

Setelah *import dataset* berhasil dilakukan, kita bisa dengan mudah mengecek informasi mengenai jumlah data dengan kelas tertentu. Untuk mengecek perbandingan jumlah data dapat kita lihat pada gambar dibawah.

Gambar 4. 4 Proses pengecekan label

Dari gambar tersebut kita mendapatkan informasi, data dengan label *hatespeech* memiliki 7574 *tweet* dengan kelas tidak *hatespeech* dan 5552 *tweet* dengan kelas *hatespeech.* Sedangkan pada label *Abusive* terdapat 8092 *tweet* dengan kelas tidak *abusive* dan 5034 *tweet* dengan kelas *Abusive*. Lalu pada label Level terdapat 3375 *tweet* dengan kelas *weak*, 1704 *tweet* dengan kelas *moderate* dan 473 *tweet* dengan kelas *strong*.

Setelah proses pengecekan, kita dapat membuang label yang tidak diperlukan lagi pada *dataset*, yaitu label HS_Weak, HS_Moderate dan HS_Strong, karena ketiga label tersebut sudah ditranformasikan kedalam label baru, yaitu label Level. Untuk membuang label yang tidak diperlukan tersebut dijelaskan digambar dibawah.

Pemisahan label

```
In [3]: data = data.drop(['HS_Weak', 'HS_Moderate', 'HS_Strong'], axis=1)

In [4]: data
```

Out[4]:

	Tweet	HS	Abusive	Level
0	- disaat semua cowok berusaha melacak perhatia...	1	1	1
1	RT USER: USER siapa yang telat ngasih tau elu?...	0	1	0
2	41. Kadang aku berfikir, kenapa aku tetap perc...	0	0	0
3	USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...	0	0	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1	1	2
...	...	...	...	...
13121	USER jangan asal ngomong ndasmu. congor lu yg ...	1	1	1
13122	USER Kasur mana enak kunyuk'	0	1	0
13123	USER Hati hati bisu :(.g\n\nlagi bosan huft \...	0	0	0
13124	USER USER USER USER Bom yang real mudah terdet...	0	0	0
13125	USER Mana situ ngasih(": itu cuma foto ya kuti...	1	1	1

13126 rows × 4 columns

Gambar 4. 5 Proses pemisahan label

Fungsi kode program *data.drop()* digunakan untuk menghapus label tertentu pada *dataset*, sehingga hasilnya *dataset* hanya tersisa tiga label yaitu HS, *Abusive* dan Level seperti yang dapat kita lihat pada gambar diatas.

Setelah label berhasil dihapus, *dataset* disimpan kedalam format csv menggunakan perintah dibawah ini.

Save Label HS Abusive dan Level

```
In [7]: data.to_csv('dataset/twitter_fix_label.csv', encoding='utf-8', index=False)
```

Gambar 4. 6 Simpan Label HS, *Abusive* dan Level

Maka dataset dengan label HS, *Abusive* dan Level seperti tercantum pada tabel berikut.

Tabel 4. 3 Dataset label HS, *Abusive* dan Level

NO	Tweet	HS	Abusive	Level
1	napadah shopee kampang'	1	1	2
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0	0	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1	0	1
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0	0	0
5	USER kolam buaya kagak ada qey'	0	0	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0	0	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--'"	0	0	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	1	1	2
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1	1	1
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0	0	0
...	...	...	...	...
13126	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0	0	0

Langkah selanjutnya adalah memisahkan data label HS, *Abusive* dengan label Level, hal ini dikarenakan jumlah data dengan label Level lebih sedikit dari data dengan label HS dan *Abusive*, yaitu data dengan label Level hanya terdiri dari data dari label HS yang bernilai benar atau 1.

```
In [12]: data = pd.read_csv('Dataset/twitter.csv')
         data = data.drop(['Level'], axis=1)

In [13]: data
```

Out[13]:

	Tweet	HS	Abusive
0	disaat semua cowok berusaha melacak perhatian...	1	1
1	RT USER USER siapa yang telat ngasih tau elu?e...	0	1
2	41 Kadang aku berfikir kenapa aku tetap percay...	0	0
3	USER USER AKU ITU AKU KU TAU MATAMU SIPIT TAPI...	0	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1	1
...	...	...	...
13121	USER jangan asal ngomong ndasmu congor lu yg s...	1	1
13122	USER Kasur mana enak kunyuk	0	1
13123	USER Hati hati bisu :(g lagi bosan huft \xf0\...	0	0
13124	USER USER USER USER Bom yang real mudah terdet...	0	0
13125	USER Mana situ ngasih(": itu cuma foto ya kuti...	1	1

13126 rows × 3 columns

Gambar 4. 7 Hapus label Level

Pada gambar diatas merupakan tahap untuk megambil data dengan label HS dan *Abusive*. Setelah berhasil *import dataset* lalu dilanjutnya dengan menghapus label Level. Setelah itu, simpan data HS dan *Abusive* seperti gambar dibawah.

Save Label HS Abusive

```
In [7]: data.to_csv('dataset/twitter_HS_Abusive.csv', encoding='utf-8', index=False)
```

Gambar 4. 8 Simpan label HS dan *Abusive*

Dataset setelah dilakukan penghapusan label Level dapat kita lihat pada tabel berikut.

Tabel 4. 4 *Dataset* Label HS dan *Abusive*

NO	Tweet	HS	Abusive
1	napadah shopee kampang'	1	1

NO	Tweet	HS	Abusive
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1	0
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0	0
5	USER kolam buaya kagak ada qey'	0	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--'"	0	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	1	1
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1	1
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0	0
...	...	...	...
13126	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0	0

Selanjutnya kita akan menyimpan *dataset* label Level, dan menghapus label HS dan *Abusive*, sehingga data dengan label Level sekarang terdiri dari 5552 *tweet.* Seperti digambar berikut.

```
In [16]: data = pd.read_csv('Dataset/twitter.csv')
         data = data[data.HS.isin([1])]
         data = data.drop(['HS', 'Abusive'], axis=1)

In [17]: data
```

Out[17]:

	Tweet	Level
0	disaat semua cowok berusaha melacak perhatian...	1
4	USER USER Kaum cebong kapir udah keliatan dong...	2
5	USER Ya bani taplak dkk \xf0\x9f\x98\x84\xf0\x...	2
10	Setidaknya gw punya jari tengah buat lu sebelo...	1
11	USER USER USER USER BANCI KALENG MALU GA BISA ...	1
...	...	...
13114	USER Pak Recep anda salah itu gubernur pakkkk ...	1
13118	brengsek itu orang terbuat dr apa bikin gue be...	1
13119	Kapolda Babi! Biadap dan Bodoh! Gak punya otak...	2
13121	USER jangan asal ngomong ndasmu congor lu yg s...	1
13125	USER Mana situ ngasih(": itu cuma foto ya kuti...	1

5552 rows × 2 columns

Gambar 4. 9 Hapus label HS dan *Abusive*

Selanjutnya kita bisa menyimpan data dengan label Level seperti digambar dibawah berikut.

Save Label Level

```
In [7]: data.to_csv('dataset/twitter_Level.csv', encoding='utf-8', index=False)
```

Gambar 4. 10 Simpan label Level

Dataset setelah dilakukan penghapusan label Level dapat kita lihat seperti pada tabel berikut.

Tabel 4. 5 *Dataset* label Level

NO	Tweet	Level
1	napadah shopee kampang'	2

NO	Tweet	Level
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0
5	USER kolam buaya kagak ada qey'	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--'"	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	2
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0
...	...	...
13126	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0

Setelah selesai pada tahap pemisahan label, tahap selanjutnya adalah melakukan *dataset splitting* atau pemisahan data. Tahap ini bertujuan untuk membagi data menjadi dua bagian, yaitu *data training* dan *data testing*. Data *training* nantinya digunakan untuk pelatihan model dan data *testing* akan diimplementasikan pada pengujian model.

```
In [30]: X_train, X_test, y_train, y_test = train_test_split(data_HS_Abusive.Tweet,
                                                            data_HS_Abusive[['HS', 'Abusive']],
                                                            test_size=0.1, shuffle=False)
```

Gambar 4. 11 *Dataset Splitting* HS dan *Abusive*

Pada gambar diatas merupakan proses untuk membagi *dataset* label HS dan *Abusive* menjadi data *training* dan data *testing*, gambar atas untuk membagi *dataset*

dengan perbandingan 90:10, sedangkan jika kita ingin membuat perbandingan 80:20 cukup dengan mengganti parameter *test_size* dengan nilai 0,2.

Setelah berhasil dibagi menjadi data *training* dan data *tesing* kita dapat menampilkan data hasil pemisahan seperti pada gambar berikut.

```
In [32]: pd.DataFrame(X_train).join(pd.DataFrame(y_train))
```

Out[32]:

	Tweet	HS	Abusive
0	disaat semua cowok berusaha melacak perhatian...	1	1
1	RT USER USER siapa yang telat ngasih tau elu?e...	0	1
2	41 Kadang aku berfikir kenapa aku tetap percay...	0	0
3	USER USER AKU ITU AKU KU TAU MATAMU SIPIT TAPI...	0	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1	1
...	...	...	...
11808	Di Maroko Anies Dipanggil â??Super Gubernurâ??	0	0
11809	sprti pilkada DKI dinyatakan Ahok elektabilita...	0	0
11810	Islam itu indah Kristen itu kasih Hindu itu ci...	0	0
11811	Perang dagang AmerikaCina lumayan mencurigakan...	0	0
11812	Kebaikan adalah bahasa yg bisa di dengar oleh ...	0	0

11813 rows × 3 columns

Gambar 4. 12 Hasil *Splitting* label HS dan *Abusive*

Selanjutnya kita bisa menyimpan *dataset training* dan *testing* untuk label HS dan *Abusive* seperti pada *script* berikut.

```
pd.DataFrame(X_train).join(pd.DataFrame(y_train)).to_csv('twitter_HS_Abusive_train.csv', encoding='utf-8', index=False)

pd.DataFrame(X_test).join(pd.DataFrame(y_test)).to_csv('twitter_HS_Abusive_test.csv', encoding='utf-8', index=False)
```

Hasil data *training* untuk label HS dan *Abusive* dapat dilihat pada tabel berikut.

Tabel 4. 6 *Dataset training* label HS dan *Abusive*

NO	Tweet	HS	Abusive
1	napadah shopee kampang'	1	1
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1	0
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0	0
5	USER kolam buaya kagak ada qey'	0	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--'"	0	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	1	1
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1	1
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0	0
...	...	...	...
1181 3	Perang dagang Amerika-Cina lumayan mencurigakan, karena merek sepatu Amerika yang "mahal" masih banyak buatan Cina.\n\nJadi? Dah beli buatan Indonesia aja \n\xf0\x9f\xa4\x99\xf0\x9f\x98\x84'	0	0

Sedangkan hasil *dataset* testing untuk label HS dan *Abusive* dapat kita dilihat pada tabel berikut.

Tabel 4. 7 *Dataset testing* label HS dan *Abusive*

NO	Tweet	HS	Abusive
1	USER USER USER USER Naik buaya sambil makan pepaya?	0	0

NO	Tweet	HS	Abusive
2	Muntahmuntah Usai Makan Seorang Banci Mengaku Telat Dua Bulan	0	0
3	Survei: Khawatir Resesi Ekonomi Konsumen Rajin Menabung	0	0
4	USER tolong si ini onta nyanyi mulu minta disemprot baygon	1	1
5	RT USER USER Tebang pilih pilih ditebang bubarkan aja USER dan bentuk lembaga baru yg benar2 independent	1	0
6	" bertudung tapi perangai tah pape " Kalau aku buat salah tegur aku Jangan babitkan tudung atau agama aku Islam tu sempurna tapi aku tak	0	0
7	Festival Kebudayaan Antarabangsa Tawau 2018 mencerminkan keistimewaan dan keunikan Sabah dari sudut kepelbagaian budaya yang ada di negeri ini ujar YAB Perdana Menteri USER ketika melancarkan festival tersebut pada malam tadi	0	0
8	USER Pak USER mudah2an Anda semakin melihat bahwa Muhammad hanya seorang buta huruf dungu yang biadab	1	1
9	USER USER Soalnya dia budek kaya bolot Om Tuuh kuping nya kotor seperti hatinya	1	1
10	USER Kamu aja mainan hole baru aku mau cari titit baru	0	1
...	...	...	...
1313	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0	0

Setelah dilakukan *splitting* pada label HS dan *Abusive* terdapat 11.813 data latih, sedangkan untuk data uji sebanyak 1.313 data untuk perbandingan pembagian data 90:10, sedangkan untuk pembagian data 80:20 terdapat 10.500 data latih dan 2.626 data uji.

Lakukan hal yang serupa untuk data dengan label Level seperti pada digambar berikut.

```
In [35]: X_train, X_test, y_train, y_test = train_test_split(data_Level.Tweet,
                                                            data_Level[['Level']],
                                                            test_size=0.1, shuffle=False)
```

Gambar 4. 13 *Dataset Splitting* Level

Setelah proses diatas selesai kita dapat menampilkan *dataset* yang telah berhasil dibagi menjadi data *training* dan data *tesing* kita dapat menampilkan data hasil pemisahan seperti digambar berikut.

```
In [36]: pd.DataFrame(X_train).join(pd.DataFrame(y_train))
```

Out[36]:

	Tweet	Level
0	disaat semua cowok berusaha melacak perhatian...	1
1	USER USER Kaum cebong kapir udah keliatan dong...	2
2	USER Ya bani taplak dkk \xf0\x9f\x98\x84\xf0\x...	2
3	Setidaknya gw punya jari tengah buat lu sebelo...	1
4	USER USER USER USER BANCI KALENG MALU GA BISA ...	1
...	...	...
4991	USER USER Bu Susi emang cuilan titit bernyawa ...	1
4992	USER tolong si ini onta nyanyi mulu minta dise...	1
4993	RT USER USER Tebang pilih pilih ditebang bubar...	3
4994	USER Pak USER mudah2an Anda semakin melihat ba...	2
4995	USER USER Soalnya dia budek kaya bolot Om Tuu...	1

4996 rows × 2 columns

Gambar 4. 14 Hasil *splitting* label Level

Untuk menyimpan data *training* dan *testing* pada label Level bisa dilakukan seperti pada *script* berikut.

```
pd.DataFrame(X_train).join(pd.DataFrame(y_train)).to_csv('twitter_Level_train.csv', encoding='utf-8', index=False)

pd.DataFrame(X_test).join(pd.DataFrame(y_test)).to_csv('twitter_Level_test.csv', encoding='utf-8', index=False)
```

Hasil data *training* untuk label Level dapat dilihat ditabel berikut.

Tabel 4. 8 *Dataset training* label Level

NO	Tweet	Level
1	USER USER USER sengkuni sekarang sudah jadi tukang ramal lho dia ramal anies penyelamat bangsa cocot e sengkuni kok dipercaya masuk neraka lo kalau percaya amin rais si tukang ramal	1
2	Sumbu pendek: jokowi koalisi dengan negara PKI Jokowi PKI lengserkan jokowi pusying hamba	3
3	USER USER Aku bukan minat pun mail Tapi looking at whole Ya dia memang dulunya cheater Tapi dia da mengaku silap dia and minta maaf Yg sampah sarap adalah kau Sshhh	1
4	USER Dulu waktu kampanye pak jokowiberkoar2 menjanjikan tarif listrik gk naikBBM gk naikmenciptakan 10000 lapangan kerja tapi nyatanyanonsens itu semua apa ini yg di bilang gk cacat?????	1
5	Ada yg bilang alexis ditutup cukup dengan surat hehe ketawain aja dah baca belum pers rilis dr manajemen alexis tp ga apa2 para kampret mah gampang dibegoin dulu jg pernah bilang ditutup eh msh exis tuh alexis	2
6	USER Apa dia pikir korban yg tewas itu bukan orang?? Manusia sontoloyo kayak gini kok bisa jadi anggota Komnas HAM?? Saya nggak rela uang pajak saya membayar gaji orang goblok yg tdk memiliki empati ini!	1
7	USER geng messi kntl tapi rapopo deng biar #Icardi &amp #Lautaro cepet nyetel	2
8	USER Kelakuan batak kafir itu ya seperti si USER Suka memecah belah umat	1
9	namanya jg simpang susun bukan bundaran semanggi Oom Habib waras? dasar katrok lo Oom	1
10	Petamina Rugi Garuda Rugi PLN rugi KAI rugi Dan banyak lagi BUMN yg harusnya untung malah devisit tapi masih ada pihak yg tanpa sungkan bilang Jokowi presiden hebat giliran anaknya bisnis pisang goreng omsetnya Triliunan mereka percaya dungu	2
...	...	...
4996	USER Terbukti kan kalo lo cuma modal fitnah??\xf0\x9f\x98\x82 Semua cebong sama modal bacot kenceng dan fitnah sana sini tanpa bisa ngebuktiin lo gak jauh beda sama temen temen lo yg di kolam \xf0\x9f\x98\x82\xf0\x9f\x98\x82	2

Hasil data *testing* label Level dapat dilihat ditabel berikut.

Tabel 4. 9 *Dataset testing* label Level

NO	Tweet	Level
1	Maju lu sini anjing	1
2	yaudah atuh usir hindu budha juga aja biar gimana pun mereka nga monoteis atau tanyain aja dulu deh tau arti kata "esa" nga sih?	3
3	RT USER USER Kata mas budi kumpul orang bodoh/gila dalam jumlah besar Murah biayanya URL	1
4	USER Baru mantap rejeki bekah buat keluarga Daripada cari sampingan jilat2 penguasa jilat2 pantat penista agama	1
5	RT USER tiap hari comment di FB pakai template yg sama: rezim PKI rezim antiislam rezim penjilat cina dsb	2
6	hih si merula tae bgt point gue dikurangi 20	1
7	Bila Cina bs maju dg Komunisme	2
8	Agus silvy Tukang Fitnah #DebatFinalPilkadaJKT	2
9	RT USER Kaum Homo Lesbian Banci kalian bela atas nama HAM tapi seorang wanita yang memakai cadar kalian cerca atas nama tol\xe2\x80\xa6	2
10	USER Salah pangkah dpt taik taik lgbtliberalmenentang raja\xc2\xb2pengampu kafir harbi	2
...	...	...
556	perempuan kaya lo mending mati aja deh jelek aja gausa sok jadi make up artist!	1

Setelah dilakukan *splitting* pada label Level terdapat 4.996 data latih dan 556 data uji untuk perbandingan pembagian data 90:10, sedangkan untuk pembagian data 80:20 terdapat 4.441 data latih dan 1.111 data uji.

4.1.2 Analisa *Training Language Model*

Model *language* yang akan dibuat menggunakan *library FastText*. Data yang akan dipergunakan untuk proses ini adalah data latih dan tanpa melibatkan data uji. Nantinya akan tercipta dua model, model pertama untuk data HS dan *Abusive* sedangkan yang kedua untuk Level.

Langkah untuk membuatnya adalah dengan melakukan *import library* yang dibutuhkan dalam proses pembuatan model. *Library* yang dibutuhkan dapat dilihat pada gambar berikut.

Import library

```
In [1]: import os
        import pandas as pd
        from tqdm.auto import tqdm
        from nltk.tokenize import word_tokenize
        from gensim.models import FastText
```

Gambar 4. 15 *Import library FastText*

Setelah itu *import dataset* yang akan digunakan untuk *training* model FastText.

Import Data

```
In [2]: df = pd.read_csv('Dataset/90-10/twitter_HS_Abusive_train.csv')
        df
```

Out[2]:

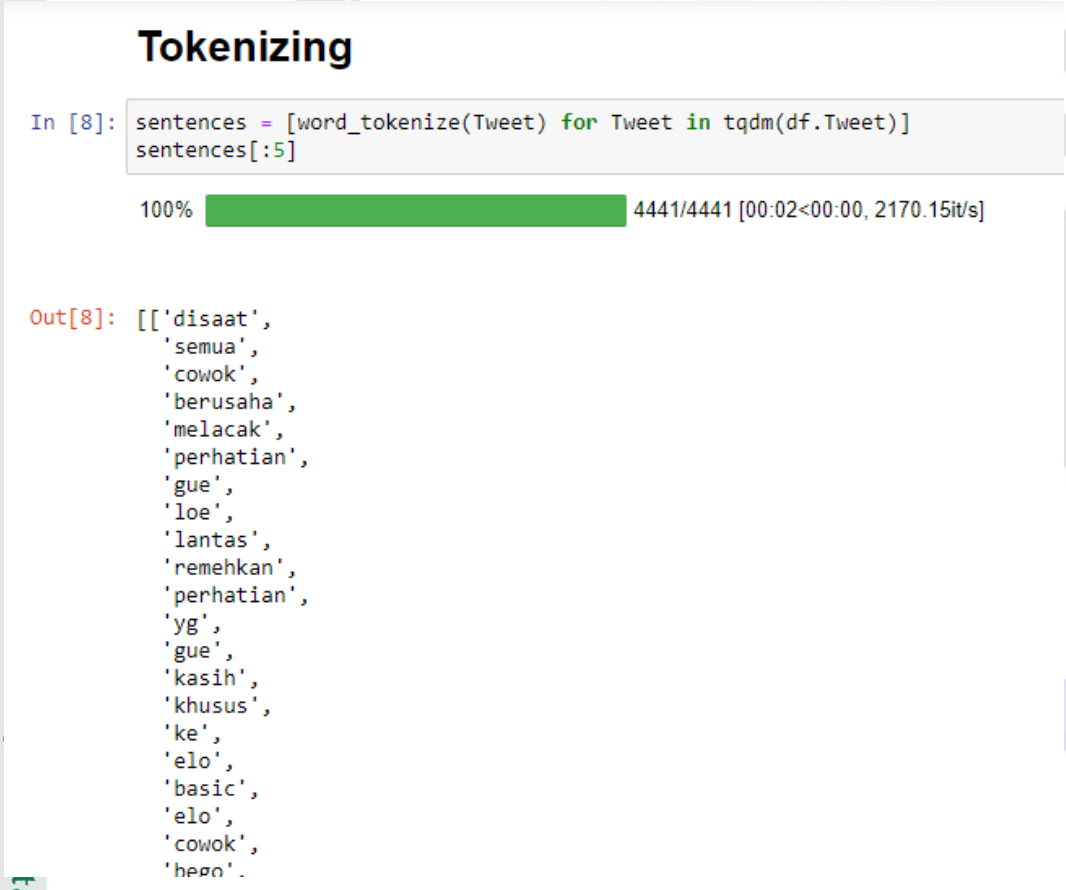
	Tweet	HS	Abusive
0	disaat semua cowok berusaha melacak perhatian...	1	1
1	RT USER USER siapa yang telat ngasih tau elu?e...	0	1
2	41 Kadang aku berfikir kenapa aku tetap percay...	0	0
3	USER USER AKU ITU AKU KU TAU MATAMU SIPIT TAPI...	0	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1	1
...	...	...	...
11808	Di Maroko Anies Dipanggil â??Super Gubernurâ??	0	0
11809	sprti pilkada DKI dinyatakan Ahok elektabilita...	0	0
11810	Islam itu indah Kristen itu kasih Hindu itu ci...	0	0
11811	Perang dagang AmerikaCina lumayan mencurigakan...	0	0
11812	Kebaikan adalah bahasa yg bisa di dengar oleh ...	0	0

11813 rows × 3 columns

Gambar 4. 16 *Import* data *training*

Dataset yang dipergunakan merupakan *dataset training* untuk label HS dan *Abusive* untuk membuat *training* model HS dan *Abusive.* Model *training* yang akan dibuat sebanyak dua, untuk perbandingan data 90:10 dan 80:20. Jumlah data training untuk perbandingan data 90:10 berjumlah 11.813 *tweet* sedangkan untuk perbandingan data 80:20 berjumlah 10.500 tweet. Sedangkan untuk membuat *training* model Level menggunakan *dataset training* Level, sama halnya dengan label HS dan *Abusive.* Model *training* yang akan dibuat sebanyak dua, untuk perbandingan data 90:10 dan 80:20. Jumlah data training untuk perbandingan data 90:10 berjumlah 4.996 *tweet* sedangkan untuk perbandingan data 4.441 berjumlah *tweet.*

Lalu data akan ditokenisasi. Hasil Tokenisasi dapat dilihat seperti pada digambar berikut.



Gambar 4. 17 Proses Tokenisasi

Selanjutnya lakukan pembuatan model *training* menggunakan data ditokenisasi sebelumnya. Prosesnya dapat dilihat digambar berikut.

Gambar 4. 18 *Training* Model

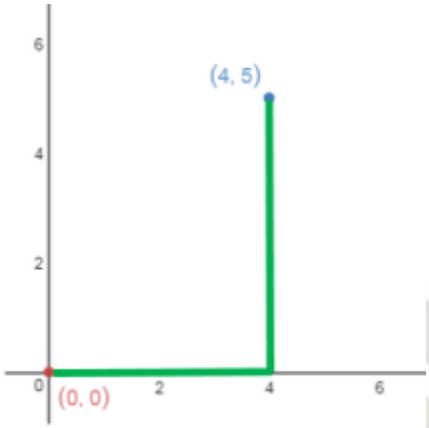
Lama proses pelatihannya berlangsung sekitar 45-60 menit. Apabila *training* selesai, kita bisa menyimpan model tersebut agar nanti bisa digunakan untuk proses klasifikasi, untuk *training* model untuk label Level dapat dilakukan dengan cara seperti diatas.

Ketika FastText model berhasil dibuat, maka dapat kita gunakan untuk proses klasifikasi. Namun sebelumnya kita perlu mengubah data berupa *word* kedalam bentuk vector. Inilah fungsi dari *model language* untuk mendapatkan *vector* dari data latih dan data uji. Setelah berhasil mendapatkan vector dari data. Kita bisa melanjutkan untuk proses klasifikasi mengunakan algoritma *Decision Tree*.

Pada proses *encoding* untuk mengubah *word vector* menjadi *sentence vector* didefenisikan dengan rumus:

$$\| x \|_1 = \sum_i | x_i |$$

Proses perhitungan ini dikenal dengan *Manhattan*. Jarak *Manhattan* tersebut mendefinisikan jumlah dari besaran *vector* disuatu ruang. Sebagai contoh diberikan *vector* $x = (4,5)$,



Gambar 4. 19 *Encoding Vector*

Maka *encoding* dari *vector* (x) ialah:

$$\| x \|_1 = |4| + |5| = 9$$

Maka *encoding* dari *vector* (x) adalah jarak yang ditempuh dari titik asal $(0,0)$ menuju $(4,5)$ (total panjang lintasan yang berwarna hijau).

4.1.3 Analisa *Text Preprocessing*

Text preprocessing dilakukan agar *dataset* yang digunakan untuk proses selanjutnya dalam keadaan bersih. Tahapan *text preprocessing* pada laporan penelitian ini diantaranya *case folding*, *cleaning, filtering* dan *tokenizing.*

Case folding digunakan untuk mengubah huruf pada *dataset tweet* menjadi huruf kecil atau *lowercase*. *Cleaning* berfungsi untuk membersihkan *dataset tweet* dari tanda baca dan simbol. *Filtering* (*stopword removal*) berfungsi untuk membuang kata umum yang biasa muncul dan tidak memiliki makna. Tahap *Tokenizing* digunakan karena data yang akan diproses kedalam model FastText berberbentuk *word* (kata), setiap *vocab* akan diubah bentuknya menjadi vektor dengan panjang 128 oleh FastText. Proses pada kode python dapat dilihat pada gambar berikut.

```
In [5]: sentences = [word_tokenize(Tweet) for Tweet in tqdm(df.Tweet)]
        sentences[:5]

        100% 11813/11813 [00:05<00:00, 2314.40it/s]

Out[5]: [['disaat',
          'semua',
          'cowok',
          'berusaha',
          'melacak',
          'perhatian',
          'gue',
          'loe',
          'lantas',
          'remehkan',
          'perhatian',
          'yg',
          'gue',
          'kasih',
          'khusus',
```

Gambar 4. 20 Tahap *preprocessing*

Kombinasi *Text Preprocessing* akan dilakukan pada *experiment* penelitan ini yang bertujuan untuk menemukan kombinasi terbaik. Selain digunakan untuk membersihkan *dataset*, variasi *text preprocessing* dilakukan untuk memilih kombinasi terbaik hal ini dikarenakan setiap variasi *text preprocessing* bisa jadi tidak untuk kondisi *dataset*, bisa jadi dapat membuat akurasi semakin naik bahkan membuat akurasi semakin turun.

Jadi, akan dilakukan *experiment* terhadap penggunaan *text preprocessing* seperti *case folding, stopword removal* dan *punctuation removal* (tanda baca). Bahkan tanpa melakukan *text preprocessing* apakah dapat meningkatkan akurasi. *Experiment* yang akan dilakukan dapat dilihat pada tabel berikut.

Tabel 4. 10 *Experiment* awal

Case Folding	***Stopword***	***Punctuation***
Ya	Ya	Ya
Ya	Ya	Tidak
Ya	Tidak	Ya
Ya	Tidak	Tidak
Tidak	Ya	Ya
Tidak	Ya	Tidak

Case Folding	*Stopword*	*Punctuation*
Tidak	Tidak	Ya
Tidak	Tidak	Tidak

4.1.4 Analisa *Feature Engineering*

Pada tahap ini akan dilakukan berbagai *experiment* untuk menemukan *set features* terbaik yang memiliki akurasi tertinggi. Untuk lebih jelas mengenai *feature engineering* yang akan dilakukan dapat dilihat ditabel berikut.

Tabel 4. 11 *Feature Engineering Set*

Kategori Fitur	Fitur	Keterangan
Khusus Twitter	F1	Jumlah tag dalam cuitan
Tekstual	F2	Jumlah kata dalam cuitan
	F3	Jumlah tanda seru dalam cutian
	F4	Jumlah tanda tanya dalam cuitan
	F5	Jumlah kata huruf kapital dalam cuitan
	F6	Jumlah kata huruf kecil dalam cuitan
Lexicon	F7	Kata berkonotasi positif
	F8	Kata berkonotasi negatif
	F9	Kata yang mengandung *vocab abusive*

Setelah dilakukan perhitungan frekuensi dari masing-masing *feature* tersebut, maka *feature* yang terbentuk digunakan sebagai penambah panjang vektor 128 yang terbentuk pada pelatihan *language* model. Adapun skema pengujian *feature engineering* yang akan dilakukan dapat dilihat pada tabel dibawah ini.

Tabel 4. 12 Kombinasi *Feature Engineering*

Percobaan ke-	Kombinasi Percobaan
1	Baseline = $[x_1, x_2, \dots, x_{128}]$
2	New Feature = $[x_1, x_2, \dots, x_{128}] \oplus [F_1]$
3	New Feature = $[x_1, x_2, \dots, x_{128}] \oplus [F_2, \dots, F_6]$

Percobaan ke-	Kombinasi Percobaan
4	New Feature = $[x_1, x_2, \dots, x_{128}] \oplus [F_7, \dots, F_9]$
5	New Feature = $[x_1, x_2, \dots, x_{128}] \oplus [F_1] \oplus [F_2, \dots, F_6]$
6	New Feature = $[x_1, x_2, \dots, x_{128}] \oplus [F_1] \oplus [F_7, \dots, F_9]$
7	New Feature = $[x_1, x_2, \dots, x_{128}] \oplus [F_2, \dots, F_6] \oplus [F_7, \dots, F_9]$
8	New Feature = $[x_1, x_2, \dots, x_{128}] \oplus [F_1] \oplus [F_2, \dots, F_6] \oplus [F_7, \dots, F_9]$

4.1.5 Analisa Klasifikasi *Decision Tree*

Klasifikasi metode *decision tree* adalah tahapan yang dilakukan untuk proses klasifikasi *hate speech, abusive* dan level pada Twitter. Untuk proses klasifikasi ini seluruh data dibagi menjadi 2 tahap, tahap pertama yaitu membagi menjadi data latih (*training*) dan tahap kedua adalah membagi menjadi data uji (*testing*). Pada penjelasan ini dilakukan perhitungan manual 7 data latih dengan panjang vektor yaitu 5, lalu model yang terbentuk nantinya akan dilakukan untuk pengujian 1 data uji.

Untuk menentukan perhitungan *splitting node* serta *terminal node, decision tree* akan menghitung nilai *gini*. Perhitungan nilai *gini* akan menentukan cabang kiri dan cabang kanan dalam pohon keputusan. Apabila perhitungan *gini* belum mencapai nilai = 0, maka *node* lanjut melakukan *splitting*. Namun, apabila nilai *gini* mencapai angka = 0, maka berhenti dalam melakukan *splitting*.

Sebagai contoh perhitungan *gini* dalam menentukan *node* paling atas menggunakan 7 buah data tweet dengan panjang vektor yaitu 5. Data yang digunakan untuk pelatihan metode *decision tree* ditampilan dibawah.

Tabel 4. 13 Data Latih

NO	Tweet	Hate Speech
1	napadah shopee kampang'	1
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0

NO	Tweet	Hate Speech
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0
5	USER kolam buaya kagak ada qey'	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--"'	0

Berikut vektor yang terbentuk dari data diatas ditampilkan ditabel berikut.

Tabel 4. 14 Vektor data latih

X[0]	X[1]	X[2]	X[3]	X[4]	Class
-1,7976288	-2,9167852	-0,17991388	-0,8077838	5,1320996	1
0,9929133	-1,6545742	-0,41695952	-1,231378	4,0469728	0
5,7353926	-1,0745741	1,1192963	2,3450077	2,665269	0
5,7673492	-1,1853626	0,18944992	0,15069261	5,716347	0
-1,3942401	-3,7312672	-0,6102253	-0,6709977	4,509192	1
-4,2454033	-2,4389486	0,1007187	-0,6343434	3,5798044	1
-2,1549864	-3,297452	1,6336719	4,5046973	-6,5897946	0

Perhitungan pertama dimulai untuk atribut X[0] dahulu, lakukan *spliting* pada atribut X[0]. *Spliting* yang memungkinkan untuk X[0] dari rentang *node* kiri, yaitu dari -4,2454033 $\leq X <$ 5,7673492. Semua nilai yang lain pada setiap *spliting* dari *node* anak untuk atribut X[0] yaitu:

1. X[0] ≤ -4,2454033
2. X[0] ≤ -2,1549864
3. X[0] ≤ -1,7976288
4. X[0] ≤ -1,3942401
5. X[0] ≤ 0,9929133
6. X[0] ≤ 5,7353926, dan
7. X[0] ≤ 5,7673492

Lalu ambil *spliting* pertama dan perhitungan *index gini* sebagai berikut.

Setelah dilakukan partisi *binary split* pada atribut X[0] ≤ -4,2454033 dengan *decision tree.*

Tabel 4. 15 *Binary Split* Pertama X[0]

Attribute	*Number of records*		
	Zero (0)	*One* (1)	N=7
X[0] ≤ -4,2454033	0	1	1
X[0] > -4,2454033	4	2	6

Kemudian hitung *gini* (d1), hitung *gini* (d2), dan *gini split* seperti berikut.

$$Gini(X[0] \leq -4{,}2454033) = 1 - (0^2 + 1^2) = 0$$

$$Gini(X[0] > -4{,}2454033) = 1 - \left(\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right) = 0{,}444444$$

$$Gini_{split} = \left(\frac{1}{7}\right) x\ 0 + \left(\frac{6}{7}\right) x 0{,}444444 = 0{,}38952$$

Pada langkah selanjutnya hitung nilai *gini* untuk X[0] ≤ -2,1549864. Setelah dilakukan partisi *binary split* pada atribut X[0] ≤ -2,1549864 dengan *decision tree*.

Tabel 4. 16 *Binary Split* Kedua X[0]

Attribute	*Number of records*		
	Zero (0)	*One* (1)	N=7
X[0] ≤ -2,1549864	1	1	2
X[0] > -2,1549864	3	2	5

Kemudian hitung *gini* (d1), hitung *gini* (d2), dan *gini split* seperti berikut.

$$Gini(X[0] \leq -2{,}1549864) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 0{,}5$$

$$Gini(X[0] > -2{,}1549864) = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right) = 0{,}48$$

$$Gini_{split} = \left(\frac{2}{7}\right) x\ 0{,}5 + \left(\frac{5}{7}\right) x\ 0{,}48 = 0{,}485714$$

Pada langkah selanjutnya hitung nilai *gini* untuk X[0] ≤ -1,7976288. Setelah dilakukan partisi *binary split* pada atribut X[0] ≤ -1,7976288 dengan *decision tree.*

Tabel 4. 17 *Binary Split* Ketiga X[0]

Attribute	*Number of records*		
	Zero (0)	*One* (1)	N=7
X[0] ≤ -1,7976288	1	2	3
X[0] > -1,7976288	3	1	4

Kemudian hitung *gini* (d1), hitung *gini* (d2), dan *gini split* seperti berikut.

$$Gini(X[0] \leq -1{,}7976288) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) = 0{,}444444$$

$$Gini(X[0] > -1{,}7976288) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = 0{,}375$$

$$Gini_{split} = \left(\frac{3}{7}\right) x\ 0{,}444444 + \left(\frac{4}{7}\right) x\ 0{,}375 = 0{,}404762$$

Pada langkah selanjutnya hitung nilai *gini* untuk X[0] ≤ -1,3942401. Setelah dilakukan partisi *binary split* pada atribut X[0] ≤ -1,3942401 dengan *decision tree.*

Tabel 4. 18 *Binary Split* Keempat X[0]

Attribute	*Number of records*		
	Zero (0)	*One* (1)	N=7
X[0] ≤ -1,3942401	1	3	4
X[0] > -1,3942401	3	0	3

Kemudian hitung *gini* (d1), hitung *gini* (d2), dan *gini split* seperti berikut.

$$Gini(X[0] \leq -1{,}3942401) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = 0{,}375$$

$$Gini(X[0] > -1{,}3942401) = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 0$$

$$Gini_{split} = \left(\frac{4}{7}\right) x\ 0{,}375 + \left(\frac{3}{7}\right) x\ 0 = 0{,}214286$$

Pada langkah selanjutnya hitung nilai *gini* untuk X[0] ≤ 0,9929133. Setelah dilakukan partisi *binary split* pada atribut X[0] ≤ 0,9929133 dengan *decision tree.*

Tabel 4. 19 *Binary Split* Kelima X[0]

Attribute	*Number of records*		
	Zero (0)	*One* (1)	N=7
X[0] ≤ 0,9929133	2	3	5
X[0] > 0,9929133	2	0	2

Kemudian hitung *gini* (d1), hitung *gini* (d2), dan *gini split* seperti berikut.

$$Gini(X[0] \leq 0{,}9929133) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 0{,}48$$

$$Gini(X[0] > 0{,}9929133) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 0$$

$$Gini_{split} = \left(\frac{5}{7}\right) x\ 0{,}48 + \left(\frac{2}{7}\right) x\ 0 = 0{,}342857$$

Pada langkah selanjutnya hitung nilai *gini* untuk X[0] ≤ 5,7353926. Setelah dilakukan partisi *binary split* pada atribut X[0] ≤ 5,7353926 dengan *decision tree.*

Tabel 4. 20 *Binary Split* Keenam X[0]

Attribute	*Number of records*		
	Zero (0)	*One* (1)	N=7
X[0] ≤ 5,7353926	3	3	6
X[0] > 5,7353926	1	0	1

Kemudian hitung *gini* (d1), hitung *gini* (d2), dan *gini split* seperti berikut.

$$Gini(X[0] \leq 5{,}7353926) = 1 - \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right) = 0{,}5$$

$$Gini(X[0] > 5{,}7353926) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 0$$

$$Gini_{split} = \left(\frac{6}{7}\right) x\ 0{,}5 + \left(\frac{1}{7}\right) x\ 0 = 0{,}428571$$

Pada langkah selanjutnya hitung nilai *gini* untuk X[0] ≤ 5,7673492. Setelah dilakukan partisi *binary split* pada atribut X[0] ≤ 5,7673492 dengan *decision tree.*

Tabel 4. 21 *Binary Split* Ketujuh X[0]

Attribute	*Number of records*		
	Zero (0)	*One* (1)	N=7
X[0] ≤ 5,7673492	4	3	7
X[0] > 5,7673492	0	0	0

Kemudian hitung *gini* (d1), hitung *gini* (d2), dan *gini split* seperti berikut.

$$Gini(X[0] \leq 5{,}7673492) = 1 - \left(\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right) = 0{,}489796$$

$$Gini(X[0] > 5{,}7673492) = 1 - \left(\left(\frac{0}{0}\right)^2 + \left(\frac{0}{0}\right)^2\right) = 0$$

$$Gini_{split} = \left(\frac{7}{7}\right) x\ 0{,}489796 + \left(\frac{0}{0}\right) x\ 0 = 0{,}489796$$

Lakukan hal yang sama seperti perhitungan diatas untuk X[1], X[2], X[3] dan X[4]. Dibawah ini merupakan hasil *gini split* untuk X[0] sebagai berikut.

Tabel 4. 22 *Gini Split* X[0]

X[0]	$Gini_{split}$
-4,2454	0,380952
-2,15499	0,485714
-1,79763	0,404762
-1,39424	0,214286
0,992913	0,342857
5,735393	0,428571
5,767349	0,489796

Perhitungan kedua kita lakukan untuk atribut X[1], lakukan *spliting* pada atribut X[1]. *Spliting* yang memungkinkan untuk X[1] dari rentang *node* kiri, yaitu dari -3,73127 ≤ X < -1,07457. Semua nilai yang lain pada setiap *spliting* dari *node* anak untuk atribut X[1] yaitu:

1. X[1] ≤ -3,73127
2. X[1] ≤ -3,29745
3. X[1] ≤ -2,91679
4. X[1] ≤ -2,43895
5. X[1] ≤ -1,65457
6. X[1] ≤ -1,18536, dan
7. X[1] ≤ -1,07457

Tabel berikut ini merupakan hasil perhitungan *gini split* untuk X[1]

Tabel 4. 23 *Gini Split* X[1]

X[1]	$Gini_{split}$
-3,73127	0,380952
-3,29745	0,485714
-2,91679	0,404762
-2,43895	0,214286
-1,65457	0,342857
-1,18536	0,428571
-1,07457	0,489796

Perhitungan ketiga kita lakukan untuk atribut X[2], lakukan *spliting* pada atribut X[2]. *Spliting* yang memungkinkan untuk X[2] dari rentang *node* kiri, yaitu dari -0,61023 ≤ X < 1,633672. Semua nilai yang lain pada setiap *spliting* dari *node* anak untuk atribut X[2] yaitu:

1. X[2] ≤ -0,61023
2. X[2] ≤ -0,41696
3. X[2] ≤ -0,17991

4. X[2] ≤ 0,100719
5. X[2] ≤ 0,18945
6. X[2] ≤ 1,119296, dan
7. X[2] ≤ 1,633672

Tabel berikut ini merupakan hasil perhitungan *gini split* untuk X[2]

Tabel 4. 24 *Gini Split* X[2]

X[2]	$Gini_{split}$
-0,61023	0,380952
-0,41696	0,485714
-0,17991	0,404762
0,100719	0,214286
0,18945	0,342857
1,119296	0,428571
1,633672	0,489796

Perhitungan keempat kita lakukan untuk atribut X[3], lakukan *spliting* pada atribut X[3]. *Spliting* yang memungkinkan untuk X[3] dari rentang *node* kiri, yaitu dari -1,231378 ≤ X < 1,633672. Semua nilai yang lain pada setiap *spliting* dari *node* anak untuk atribut X[3] yaitu:

1. X[3] ≤ -1,231378
2. X[3] ≤ -0,8077838
3. X[3] ≤ -0,67099774
4. X[3] ≤ -0,6343434
5. X[3] ≤ 0,15069261
6. X[3] ≤ 2,3450077, dan
7. X[3] ≤ 4,5046973

Tabel berikut ini merupakan hasil perhitungan gini split untuk X[3]

Tabel 4. 25 *Gini Split* X[3]

X[3]	$Gini_{split}$
-1,231378	0,428571
-0,8077838	0,485714
-0,67099774	0,404762
-0,6343434	0,214286
0,15069261	0,342857
2,3450077	0,428571
4,5046973	0,489796

Perhitungan kelima kita lakukan untuk atribut X[4], lakukan *spliting* pada atribut X[4]. *Spliting* yang memungkinkan untuk X[4] dari rentang *node* kiri, yaitu dari -6,58979 ≤ X < 5,716347. Semua nilai yang lain pada setiap *spliting* dari *node* anak untuk atribut X[4] yaitu:

1. X[4] ≤ -6,58979
2. X[4] ≤ 2,665269
3. X[4] ≤ 3,579804
4. X[4] ≤ 4,046973
5. X[4] ≤ 4,509192
6. X[4] ≤ 5,1321, dan
7. X[4] ≤ 5,716347

Tabel berikut ini merupakan hasil perhitungan gini split untuk X[4]

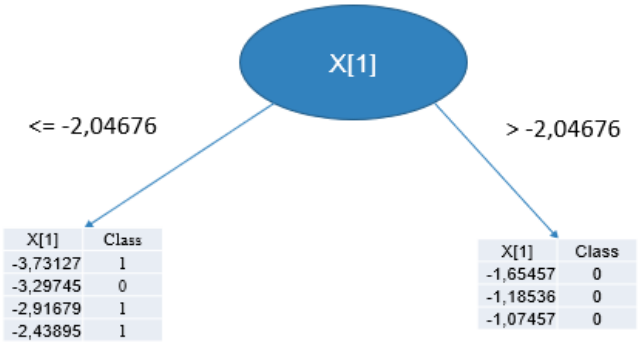
Tabel 4. 26 *Gini Split* X[4]

X[4]	$Gini_{split}$
-6,58979	0,428571
2,665269	0,342857
3,579804	0,47619
4,046973	0,404762

X[4]	$Gini_{split}$
4,509192	0,485714
5,1321	0,428571
5,716347	0,489796

Dari tabel $Gini_{split}$ diatas, didapat nilai terendah pada X[0], X[1] dan X[2] dengan nilai 0,214286. Maka nilai yang diambil boleh *random*, disini kita ambil nilai X[1]. Tetapi karena nilai pada atribut X[1] yang bersifat kontinyu, maka titik tengah dari setiap pasangan nilai yang berturut-turut terpilih sebagai titik *best split*. Sehingga nilai *best split* untuk X[1] adalah $\frac{(-2,43895)+(-1,65457)}{2} = -2,04676$

Decision tree setelah *split* pertama ditampilkan dalam gambar dibawah ini.



Gambar 4. 21 *Decision Tree Split* Pertama

Langkah selanjutnya tentukan data untuk cabang kanan dan cabang kiri, disini kita akan membandingkan data X[1] dengan nilai -2,04676, apabila nilai X[1] ≤ -2,04676 maka data tersebut masuk kedalam cabang kiri, namun apabila nilai X[1] > -2,04676, maka masuk kedalam cabang kanan, berikut hasil perbandingannya.

Tabel 4. 27 Perbandingan nilai *split* X[1]

X[0]	X[1]	X[2]	X[3]	Class	Cabang
-1,39424	-3,73127	-0,61023	-0,671	1	Kiri
-2,15499	-3,29745	1,633672	4,504697	0	Kiri
-1,79763	-2,91679	-0,17991	-0,80778	1	Kiri
-4,2454	-2,43895	0,100719	-0,63434	1	Kiri

X[0]	X[1]	X[2]	X[3]	Class	Cabang
0,992913	-1,65457	-0,41696	-1,23138	0	Kanan
5,767349	-1,18536	0,18945	0,150693	0	Kanan
5,735393	-1,07457	1,119296	2,345008	0	Kanan

Karena cabang kanan semua kelas pada datanya adalah 0, maka cabang kanan merupakan daun dari pohon, namun pada cabang kiri datanya masih bervariasi maka kita cari *gini split* seperti yang pertama kali kita lakukan, *gini split* yang kita lakukan selanjutnya adalah untuk data X[2] cabang kiri.

Perhitungan untuk cabang kedua kita lakukan untuk atribut X[2] cabang kiri, lakukan *spliting* pada atribut X[2] cabang kiri. *Spliting* yang memungkinkan untuk X[2] cabang kiri dari rentang *node* kiri, yaitu dari -0,61023 $\leq X <$ 1,633672. Semua nilai yang lain pada setiap *spliting* dari *node* anak untuk atribut X[2] yaitu:

1. X[2] ≤ -0,61023
2. X[2] ≤ -0,17991
3. X[2] ≤ 0,100719
4. X[2] ≤ 1,633672

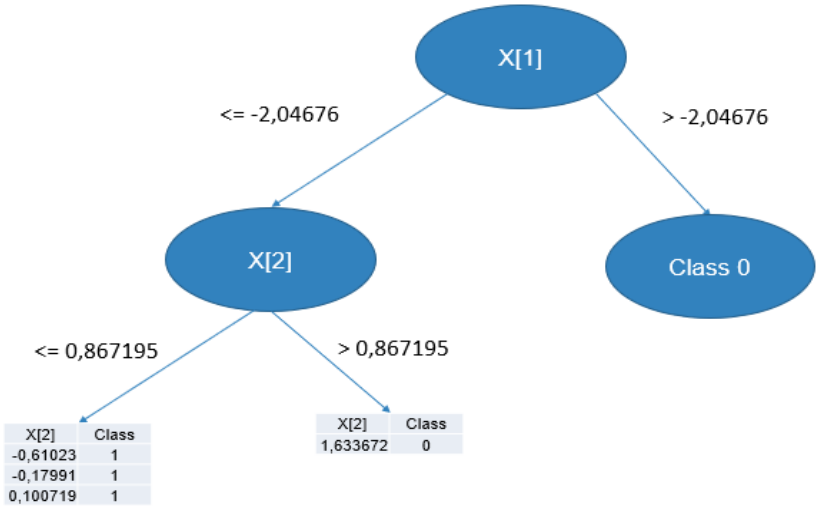
Tabel berikut ini merupakan hasil perhitungan *gini split* untuk X[2]

Tabel 4. 28 *Gini Split* X[2] Cabang Kiri

X[2]	$Gini_{split}$
-0,61023	0,333333
-0,17991	0,25
0,100719	0
1,633672	0

Dari tabel $Gini_{split}$ diatas, didapat nilai terendah pada X[2] dengan nilai 0,100719. Tetapi karena nilai pada atribut X[2] yang bersifat kontinyu, maka titik tengah dari setiap pasangan nilai yang berturut-turut terpilih sebagai titik *best split*. Sehingga nilai *best split* untuk X[2] adalah $\frac{(0,100719)+(1,633672)}{2} = 0,867195$

Decision tree setelah *split* kedua ditampilkan dalam gambar dibawah ini.



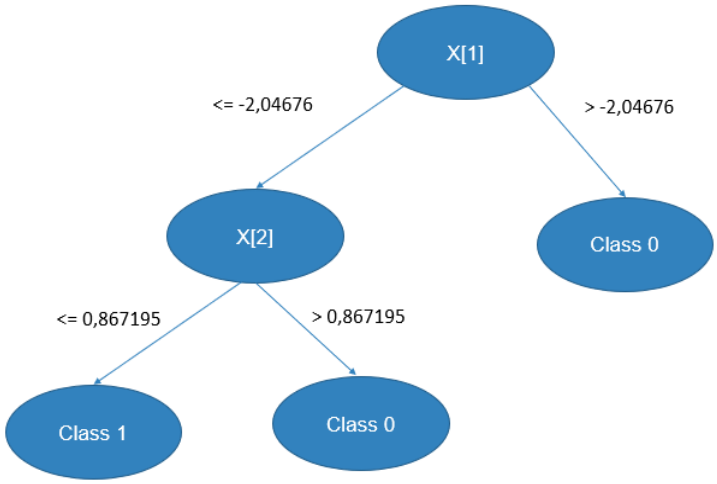
Gambar 4. 22 *Decision Tree Split* Kedua

Langkah selanjutnya tentukan data untuk cabang kanan dan cabang kiri, disini kita akan membandingkan data X[2] dengan nilai 0,867195, apabila nilai X[2] ≤ 0,867195 maka data tersebut masuk kedalam cabang kiri, namun apabila nilai X[2] > 0,867195, maka masuk kedalam cabang kanan, berikut hasil perbandingannya.

Tabel 4. 29 Perbandingan nilai *split* X[2]

X[2]	Class	Cabang
-0,61023	1	Kiri
-0,17991	1	Kiri
0,100719	1	Kiri
1,633672	0	Kanan

Karena cabang kanan semua kelas pada datanya adalah 0, maka cabang kanan merupakan daun dari pohon, begitu juga dengan cabang kiri, semua kelas pada datanya adalah 1, maka cabang kiri juga merupakan daun dari pohon keputusan yang terbentuk. Berikut ini *Decision tree* setelah *split* kedua ditampilkan dalam gambar dibawah ini.



Gambar 4. 23 *Decision Tree* Akhir

Berdasarkan pohon keputusan yang terbentuk pada gambar diatas, jika ada data baru maka data tersebut akan diklasifikasikan berdasarkan *rule* pada pohon keputusan yang terbentuk. Berikut contoh data untuk pengujian yang dilakukan.

Tabel 4. 30 Data Uji Manual

NO	Tweet	HS
1	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	?

Nilai vektor yang terbentuk ditampilkan ditabel berikut.

Tabel 4. 31 Vektor data uji manual

X[0]	X[1]	X[2]	X[3]	X[4]	HS
-4,981685	-1,535238	-0,905815	-3,367085	3,203572	?

Pertanyaannya, apakah data tersebut masuk kedalam *hate speech* atau tidak *hate speech*, maka berdasarkan model pohon *Decision tree* yang terbentuk sebelumnya maka, data tweet tersebut masuk kedalam kelas 0 atau tidak *hate speech*, hal ini karena nilai X[1] > -2,04676.

Tabel 4. 32 Hasil Klasifikasi manual

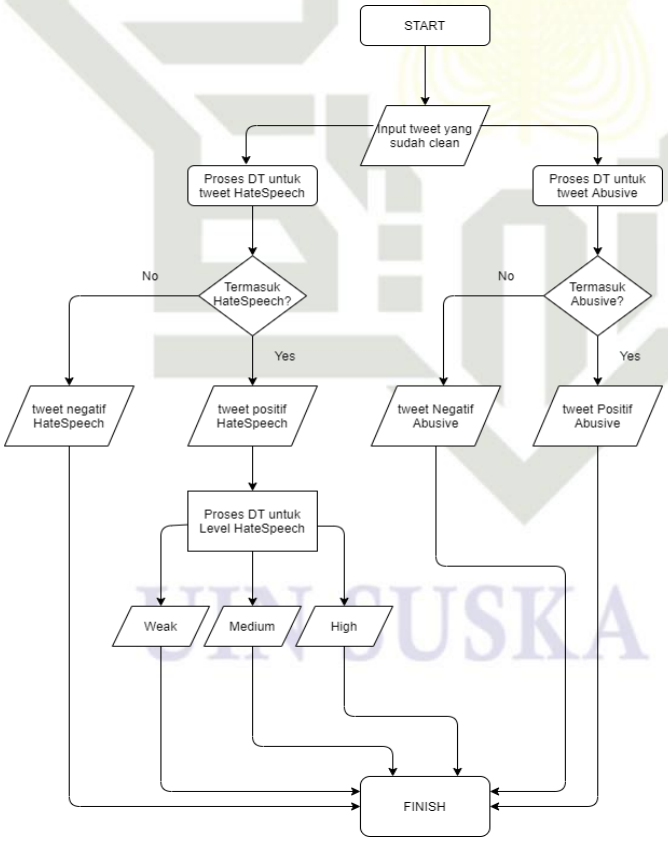
NO	Tweet	HS
1	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0

4.2 Perancangan

Pada tahapan ini dijelaskan bagaimana proses perancangan model klasifikasi menggunakan metode decision tree.

4.2.1 *Decision Tree*

Klasifikasi yang dilakukan menggunakan algoritma *Decision Tree* pada penelitian ini digambarkan pada *flowchart* dibawah ini.

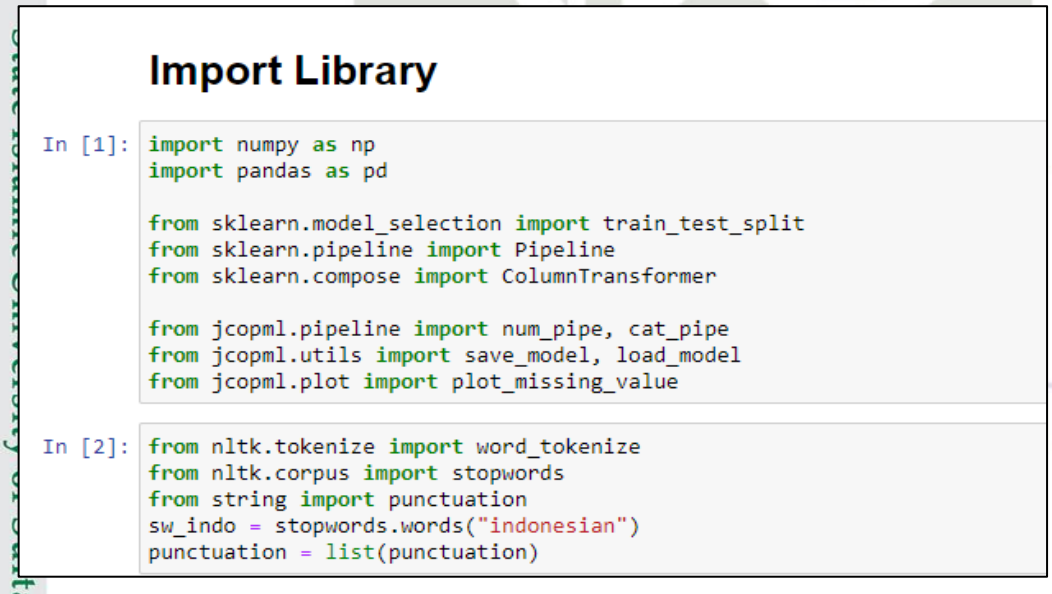


Gambar 4. 24 Alur klasifikasi *Decision Tree*

Untuk proses klasifikasi penelitian ini menggunakan algoritma *Decision Tree* dimulai dari data awal berupa *tweet. Tweet* yang diinputkan akan dilakukan pengecekan apakah *tweet* tersebut masuk kedalam kelas *hate speech* dan *abusive*. Jika *tweet* tersebut masuk kedalam kelas tidak *hate speech* (tidak mengandung *hate speech*) maka proses pengecekan untuk label *hate speech* selesai dan menghasilkan *tweet* bernilai negatif (tidak mengandung *hate speech*), namun apabila *tweet* tersebut masuk kedalam kelas *hate speech* (mengandung *hate speech*) maka dilakukan pengecekan label selanjutnya, yaitu *hate speech* level dari *tweet* tersebut, termasuk kedalam level mana, antara lemah, sedang ataupun tinggi.

Selanjutnya *tweet* tersebut akan dilakukan klasifikasi terhadap label *abusive*, apakah *tweet* tersebut masuk kedalam kelas *abusive* (mengandung *abusive*) atau tidak masuk kedalam kelas *abusive* (tidak mengandung *abusive*). Setelah proses selesai, maka mesin berhenti dalam melakukan klasifikasi terhadap *tweet.*

Untuk implementasi *decision tree* pada penelitian ini akan menggunakan datas berupa vector dari kalimat pada dataset. Library yang digunakan dapat dilihat digambar berikut.



Gambar 4. 25 *Import Library*

Lanjutkan proses mengimport data latih dan uji ketiga label yang telah diproses.

Import Data

```
In [3]: bagi = '90-10'
        df_train = pd.read_csv('Dataset/'+bagi+'/twitter_HS_Abusive_train.csv')
        df_test = pd.read_csv('Dataset/'+bagi+'/twitter_HS_Abusive_test.csv')
        df_level_train = pd.read_csv('Dataset/'+bagi+'/twitter_Level_train.csv')
        df_level_test = pd.read_csv('Dataset/'+bagi+'/twitter_Level_test.csv')
```

Gambar 4. 26 *Import Data*

Lalu *load* model pelatihan yang telah disimpan. Model ini digunakan untuk tahap klasifikasi.

Load Model FastText

```
In [4]: from gensim.models import FastText

In [5]: w2v = FastText.load('model/bismillah/'+bagi+'/twitter_HS_Abusive.fasttext').wv
        w2v_level = FastText.load('model/bismillah/'+bagi+'/twitter_Level.fasttext').wv
```

Gambar 4. 27 *Load FastText Model*

Sebelum melanjutkan ke proses klasifikasi, kita harus menjumlahkan data vektor tersebut terlebih dahulu, karena data yang kita miliki saat ini masih berupa vektor kata. Untuk menggunakannya pada tahap klasifikasi, vektor harus diubah menjadi vektor kalimat (kalimat). Karena digunakan dalam bentuk *sentence* (kalimat) saat menguji input data. Kodingan untuk proses *control vocab* kalimat dapat dilihat digambar berikut.

Control Vocab

```
In [6]: def sent_vector(sentence, w2v, stopwords=None):
            if stopwords is None:
                vecs = [w2v[word.lower()] for word in word_tokenize(sentence)]
            else:
                vecs = [w2v[word.lower()] for word in word_tokenize(sentence) if word not in stopwords]
            sent_vec = np.mean(vecs, axis=0)
            return sent_vec
```

Gambar 4. 28 *Control Vocab*

Apabila vektor kalimat berhasil didapatkan , maka *sentence vector* tersebut digunakan dalam proses pelatihan antara data, model *FastText*, dan *preprocessing* teks. Untuk proses pelatihan, Label HS dan *Abusive* menggunakan model pelatihan bernama "w2v", dan untuk label Level menggunakan model pelatihan bernama "w2v_level".

Dalam proses pelatihan ini, eksperimen dapat dilakukan dengan menggabungkan proses *preprocessing* teks dengan model pelatihan, vektor kalimat, data latih, dan pengujian setiap label.

Berikut kode program untuk mendapatkan vector kalimat dengan menggunakan *stopword removal* dan *punctuation removal* pada label HS dan *Abusive.*

Vektor Data Latih HS Abusive

```
In [7]: pakai_sw = sw_indo+punctuation

In [8]: vecs_train = [sent_vector(sentence, w2v, stopwords=pakai_sw) for sentence in df_train.Tweet]
        vecs_train = np.array(vecs_train)
        vecs_train.shape

Out[8]: (11813, 128)
```

Vektor Data Test HS Abusive

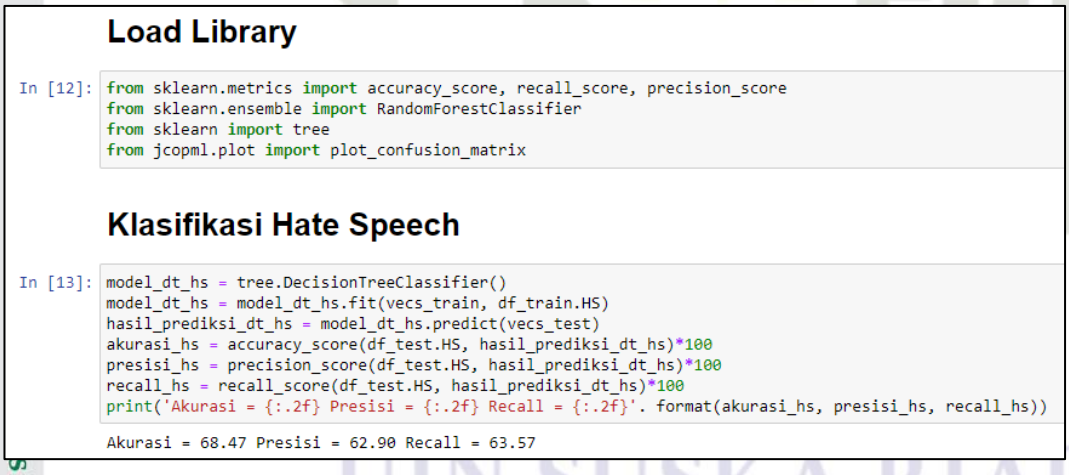
```
In [9]: vecs_test = [sent_vector(sentence, w2v, stopwords=pakai_sw) for sentence in df_test.Tweet]
        vecs_test = np.array(vecs_test)
        vecs_test.shape

Out[9]: (1313, 128)
```

Gambar 4. 29 Kode Program Vektor latih dan uji label HS dan *Abusive*

Berikut kode program untuk mendapatkan vector kalimat dengan menggunakan *stopword removal* dan *punctuation removal* pada label Level.

Gambar 4. 30 Kode Program Vektor latih dan uji label Level

Setelah menyelesaikan pelatihan tentang ketiga label tersebut. Akan kita mulai proses klasifikasinya menggunakan *decision tree*. Sebelumnya, hal pertama kita lakukan adalah *import library* untuk menghitung nilai akurasi dan klasifikasi decision treenya. Berdasarkan jumlah label yang ada, proses klasifikasi dilakukan sebanyak tiga kali. Setelah setiap label mencapai akurasinya, itu ditambahkan lalu dibagi dengan banyaknya label untuk mendapatkan nilai akurasi rata-rata. Kode program yang digunakan untuk klasifikasi decision tree pada label HS ditunjukkan pada Gambar di bawah ini.



Gambar 4. 31 Klasifikasi *Decision Tree*

Kode program yang digunakan untuk klasifikasi *decision tree* pada label *Abusive* ditunjukkan digambar berikut.

Klasifikasi Abusive

```
In [15]: model_dt_abusive = tree.DecisionTreeClassifier()
         model_dt_abusive = model_dt_abusive.fit(vecs_train, df_train.Abusive)
         hasil_prediksi_dt_abusive = model_dt_abusive.predict(vecs_test)
         akurasi_abusive = accuracy_score(df_test.Abusive, hasil_prediksi_dt_abusive)*100
         presisi_abusive = precision_score(df_test.Abusive, hasil_prediksi_dt_abusive)*100
         recall_abusive = recall_score(df_test.Abusive, hasil_prediksi_dt_abusive)*100
         print('Akurasi = {:.2f} Presisi = {:.2f} Recall = {:.2f}'. format(akurasi_abusive, presisi_abusive, recall_abusive))

         Akurasi = 71.90 Presisi = 61.28 Recall = 63.69
```

Gambar 4. 32 Klasifikasi *Abusive*

Kode program yang digunakan untuk klasifikasi *decision tree* pada label *Level* ditunjukkan digambar berikut.

Klasifikasi Level

```
In [16]: model_dt_level = tree.DecisionTreeClassifier()
         model_dt_level = model_dt_level.fit(vecs_level_train, df_level_train.Level)
         hasil_prediksi_dt_level = model_dt_level.predict(vecs_level_test)
         akurasi_level = accuracy_score(df_level_test.Level, hasil_prediksi_dt_level)*100
         presisi_level = precision_score(df_level_test.Level, hasil_prediksi_dt_level, average='micro')*100
         recall_level = recall_score(df_level_test.Level, hasil_prediksi_dt_level, average='micro')*100
         print('Akurasi = {:.2f} Presisi = {:.2f} Recall = {:.2f}'.format(akurasi_level, presisi_level, recall_level))

         Akurasi = 58.45 Presisi = 58.45 Recall = 58.45
```

Gambar 4. 33 Klasifikasi Level

BAB VI
PENUTUP

6.1 Kesimpulan

Kesimpulan yang dapat ditarik berdasarkan hasil implementasi dan pengujian yang telah dilakukan adalah sebagai berikut:

1. Metode *Decision Tree* dapat diterapkan untuk proses klasifikasi *multi-class* dan *multi-label hate speech* dan *abusive language* pada Twitter berbahasa Indonesia.
2. Kombinasi *Preprocessing* terbaik tanpa *feature engineering,* adalah menggunakan *Case Folding*, tanpa *Stopword Removal* dan tanpa *Punctuation Removal* dengan rata-rata nilai akurasi sebesar **69,77%**.
3. Penggunaan *feature engineering* dapat meningkatkan akurasi, akurasi rata-rata tertinggi didapat pada *New Feature* (8) atau *baseline* + Fitur Khusus + Fitur Tekstual + Fitur *Lexicon* dengan nilai rata-rata akurasi sebesar **71,03%**.
4. Model *Hate Speech* terbaik adalah *New feature* (2) tanpa kombinasi *preprocessing* dengan nilai akurasi sebesar **72,20%**.
5. Model *Abusive* terbaik adalah *New Feature* (8) dengan penggunaan *preprocessing Case Folding* dengan nilai akurasi sebesar **78,22%**.
6. Model Level terbaik adalah *Baseline* (1) dengan kombinasi *preprocessing Case Folding* + *Stopword Removal* + *Punctuation Removal* dengan nilai akurasi sebesar **64,39%**.

6.2 Saran

Saran yang dapat diberikan untuk penelitian selanjutnya yang terkait dengan penelitian ini adalah sebagai berikut:

1. Klasifikasi selanjutnya dapat dilakukan dengan menggunakan teknik *word embedding* lainnya seperti GloVe atau Word2Vec untuk menentukan *word embedding* yang paling baik.
2. Pada penelitan selanjutnya dapat diterapkan menggunakan *tweet* bahasa daerah.
3. Pada penelitian selanjutnya dapat menggunakan algoritma klasifikasi lainnya seperti *Naïve Bayes Classifier*, *Support Vector Machine* atau bahkan menggunakan metode *deep learning* seperti *Convolutional Neural Network* dan *Recurrent Neural Network.*

DAFTAR PUSTAKA

Alfariqi, F., Maharani, W., & Husen, J. H. (2020). *Klasifikasi Sentimen pada Twitter dalam Membantu Pemilihan Kandidat Karyawan dengan Menggunakan Convolutional Neural Network dan Fasttext Embeddings Pendahuluan*. *7*(2), 8052–8062.

Ariestya, W. W., Praptiningsih, Y. E., & Supriatin, W. (2016). Decision Tree Learning Untuk Penentuan Jalur Kelulusan Mahasiswa. *Jurnal Ilmiah FIFO*, *8*(1), 97. https://doi.org/10.22441/fifo.v8i1.1304

Arliyanti Nurdin, Bernadus Anggo Seno Aji, Anugrayani Bustamin, Z. A. (2020). Perbandingan kinerja Word Embedding Word2Vec, Glove dan FastText pada klasifikasi teks. *Jurnal TEKNOKOMPAK*, *14*(2), 74--79.

Baihaqi, D. I., Handayani, A. N., & Pujianto, U. (2019). Perbandingan Metode Naïve Bayes Dan C4.5 Untuk Memprediksi Mortalitas Pada Peternakan Ayam Broiler. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *10*(1), 383–390. https://doi.org/10.24176/simet.v10i1.2846

Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal Maret*, *1*(1), 32–41. Retrieved from https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language ∗*.

Febriyani, M. (2018). Analisis faktor penyebab pelaku melakukan ujaran kebencian (hate speech) dalam media sosial. *Journal of Linguistics*, *3*(2), 139–157. https://doi.org/10.18041/2382-3240/saber.2010v5n1.2536

Feldman, R. (2008). *Text Mining and Link Analysis*.

Garla, V. N., & Brandt, C. (2012). Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics*, *45*(5), 992–998. https://doi.org/10.1016/j.jbi.2012.04.010

Hakiem, M., & Fauzi, M. A. (2019). *Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain*. *3*(3), 2443–2451.

Hastuti, K. (2012). Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif. *Semantik*.

Hidayatullah, A. F., Yusuf, A. A. F., Juwairi, K. P., & Nayoan, R. A. N. (2019). Identifikasi Konten Kasar pada Tweet Bahasa Indonesia. *Jurnal Linguistik Komputasional*, *2*(1).

Ibrohim, M. O., & Budi, I. (2019). *Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter*. 46–57. https://doi.org/10.18653/v1/w19-3506

Komnas HAM. (2015). Buku Saku Penanganan Ujaran Kebencian (Hate Speech). *Jakarta: KOMNAS HAM*, 10.

Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). *Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine*. *2*(11), 4704–4713.

Nugroho, M. F., & Wibowo, S. (2017). *Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes*. *3*(1), 63–70.

Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan

Jumlah Retweet. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *1*(12), 1750–1757. https://doi.org/10.1074/jbc.M209498200

Putra, E. D. (2014). *Menguak Jejaring Sosial*. Tangerang.

Romadloni, N. T., Santoso, I., & Budilaksono, S. (2019). *Perbandingan metode naive bayes, knn dan decision tree terhadap analisis sentimen transportasi commuter line*. *3*(2), 1–9.

Setiawati, D., Taufik, I., & Z, W. B. (2016). *KLASIFIKASI TERJEMAHAN AYAT AL-QURAN TENTANG ILMU SAINS MENGGUNAKAN ALGORITMA DECISION TREE BERBASIS MOBILE*. *I*(1), 24–27.

Srivastava, A. N., & Sahami, M. (2009). *Text mining Classification, clustering, and applications*. https://doi.org/https://doi.org/10.1017/CBO9781107415324.004

LAMPIRAN A

DATASET AWAL DENGAN LABEL

Berikut ini merupakan dataset awal berserta label dan kelas yang digunakan pada penelitian ini:

NO	Tweet	Hate Speech	Abusive	Hate Speech Level		
				Weak	Moderate	Strong
1	napadah shopee kampang'	1	1	1	0	0
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0	0	0	1	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1	0	0	0	0
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0	0	0	0	0
5	USER kolam buaya kagak ada qey'	0	0	0	0	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0	0	0	0	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--"'	0	0	0	0	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	1	1	0	1	0
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1	1	1	0	0
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0	0	0	0	0
11	ini jalan dua platform ya di fb juga jalan bom yang gede klo mau dapet	0	0	0	0	0

NO	Tweet	Hate Speech	Abusive	Hate Speech Level		
				Weak	Moderate	Strong
12	Dikira org katolik coba:(	0	0	0	0	0
13	USER berarti lo cacat	1	1	1	0	0
14	USER USER USER USER USER Hahahaha roti buaya mah buat nikahan pak	0	0	0	0	0
15	USER USER luhut bajingan	1	1	1	0	0
16	Me every night: AAAAA ANJENG GUE MALU BANGSAD MAU ILANG AJA AAAAAAAAA APAAN SIH GUE JIJIK BANGET AAAAAAAAA INI APAAN SIH GUE DULU NGGAK DANTA AAAAAAAAAAA MAU ILANG AJA AAAAAAAAAAAAAAA	0	1	0	0	0
17	USER Terus?? Alus nya di puji Butut nya disebut butut Goblog disebut goblog	0	1	0	0	0
18	Yang sabar gua mah punya adek2 kunyuk yg bau kek sempak kuda	0	1	0	0	0
19	Tuhan itu iBlis Iblis itu Tuhan Agama onta	1	1	0	1	0
20	USER USER USER USER USER USER USER USER Assalamu alaikm ibu USER tlng hrp bantuanx u kmi #korbanphkfreeport	0	0	0	0	0
...	...	...	...	...	...	...
1312 6	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0	0	0	0	0

LAMPIRAN B

DATASET SETELAH PENGHAPUSAN LABEL

Berikut ini merupakan hasil penghapusan label dimana label yang tersisa adalah label HS dan *Abusive*:

NO	Tweet	HS	Abusive
1	napadah shopee kampang'	1	1
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1	0
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0	0
5	USER kolam buaya kagak ada qey'	0	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--'"	0	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	1	1
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1	1
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0	0
11	ini jalan dua platform ya di fb juga jalan bom yang gede klo mau dapet	0	0
12	Dikira org katolik coba:(	0	0
13	USER berarti lo cacat	1	1
14	USER USER USER USER USER Hahahaha roti buaya mah buat nikahan pak	0	0
15	USER USER luhut bajingan	1	1

NO	Tweet	HS	Abusive
16	Me every night: AAAAA ANJENG GUE MALU BANGSAD MAU ILANG AJA AAAAAAAAA APAAN SIH GUE JIJIK BANGET AAAAAAAAA INI APAAN SIH GUE DULU NGGAK DANTA AAAAAAAAAAA MAU ILANG AJA AAAAAAAAAAAAAA	0	1
17	USER Terus?? Alus nya di puji Butut nya disebut butut Goblog disebut goblog	0	1
18	Yang sabar gua mah punya adek2 kunyuk yg bau kek sempak kuda	0	1
19	Tuhan itu iBlis Iblis itu Tuhan Agama onta	1	1
20	USER USER USER USER USER USER USER USER Assalamu alaikm ibu USER tlng hrp bantuanx u kmi #korbanphkfreeport	0	0
...	...	...	...
13126	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0	0

Berikut ini merupakan hasil penghapusan label dimana label yang tersisa adalah label Level.

NO	Tweet	Level
1	napadah shopee kampang'	2
2	Bu, di perpustakaan orangnya pada bisu ya bu? | hah? Emang knp nak? | tuh mereka pada diem2an, ga ada yg bersuara | -__- ,'	0
3	lu baca bandingkan modal dgn hasil diperpanjang atau tidak dasar antek asing	1
4	USER Kerjaan? Tugas prakarya seni budaya? :3	0
5	USER kolam buaya kagak ada qey'	0
6	USER Yu...mari kita songsong presiden baru ditahun 2019	0
7	USER USER Mana mungkin w bani ngepost milik orang lain nik--'"	0
8	USER USER Cebong tidak akan pernah senang lihat kejayaan pk anis mereka akan ttp sirik. Yg parahnya klu junjungan dipuji habis habisan meskipun melakukan hal buruk.'	2

NO	Tweet	Level
9	RT USER: USER Klo busuk mawati beragama islam .. Maka dia Takut dg dirinya sendiri\nGejala gila nich nenek ..'	1
10	Salah satu obyek vital yg pengamanannya lemah itu gedung DPR. Smg dgn disahkannya UU Antiterorisme yg baru, DPR tak jadi sasaran bom bunuh diri. Amin. *Tapi doa saya biasanya tak terkabul \xf0\x9f\x98\x89*'	0
11	ini jalan dua platform ya di fb juga jalan bom yang gede klo mau dapet	0
12	Dikira org katolik coba:(	0
13	USER berarti lo cacat	0
14	USER USER USER USER USER Hahahaha roti buaya mah buat nikahan pak	1
15	USER USER luhut bajingan	0
16	Me every night: AAAAA ANJENG GUE MALU BANGSAD MAU ILANG AJA AAAAAAAAA APAAN SIH GUE JIJIK BANGET AAAAAAAAA INI APAAN SIH GUE DULU NGGAK DANTA AAAAAAAAAAA MAU ILANG AJA AAAAAAAAAAAAAAA	1
17	USER Terus?? Alus nya di puji Butut nya disebut butut Goblog disebut goblog	0
18	Yang sabar gua mah punya adek2 kunyuk yg bau kek sempak kuda	0
19	Tuhan itu iBlis Iblis itu Tuhan Agama onta	0
20	USER USER USER USER USER USER USER USER Assalamu alaikm ibu USER tlng hrp bantuanx u kmi #korbanphkfreeport	2
...	...	...
1312 6	USER USER USER USER Bom yang real mudah terdeteksi bom yang terkubur suatu saat lebih dahsyat ledakannya itulah di sebut Revolusi Jiwa'	0

DAFTAR RIWAYAT HIDUP

Nama Lengkap : Fauzi Ihsan

Tempat/Tanggal Lahir : Muara Bungkal, 13-02-1997

Agama : Islam

Anak Ke : 3

Jumlah Saudara : 6

Alamat : Jl. Uka, Perumahan Nugraha Permata, Blok J No 9

Email : 11651100236@students.uin-suska.ac.id

PENDIDIKAN

2004-2010 : SDN 05 Muara Bungkal

2010-2013 : SMPN 2 Satu Atap Sungai Mandau

2013-2016 : SMAN 1 Sungai Mandau

2016-2021 : Universitas Islam Negeri Sultan Syarif Kasim Riau, Jurusan Teknik Informatika

PENGALAMAN ORGANISASI

2014-2015 : Ketua MPK SMAN 1 Sungai Mandau

2016-2020 : Penerima Beasiswa Perguruan Tinggi *Community Development* PT.RAPP